UNIVERSITY OF
PLYMOUTH

AN INVESTIGATION INTO THE USE OF ARTIFICIAL INTELLIGENCE TECHNIQUES FOR THE ANALYSIS AND CONTROL OF INSTRUMENTAL TIMBRE AND TIMBRAL COMBINATIONS

by

AURÉLIEN ANTOINE

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Humanities and Performing Arts

July 2018

This thesis is dedicated to my Dad, Philippe Antoine, who sadly passed away during my time as a Ph.D. candidate...

Acknowledgements

First, I would like to express my gratitude to my Director of Study, Professor Eduardo R. Miranda, for believing in me and for his continuous guidance and advice during this research. I would also like to thank my supervisors Dr Alexis Kirke and Dr Marcelo Gimenes for their support and encouragement.

Next, it is important to acknowledge and thank the AHRC-funded 3D3 Centre for Doctoral Training for awarding me a PhD studentship. Without it, this research would have not been possible. I would also like to thank all the 3D3 people I had the chance to meet for the conversations and activities we shared.

A special thank to my awesome friends Ed, Jared, and Ben. Thank you for all the laughs, chats, and activities, for your help and support any time I needed it, and for the lunches at the table, the perfect break during the long working days. You were my family on the other side of the Channel.

Many thanks to all my fellow colleagues at the ICCMR, Joel, Duncan, Rodrigo, Federico, Pierre-Emmanuel, Nuria, Satvik, Michael, and Richard for the discussions, ideas, laughs, and support during the many hours spent in the lab. Thank you also for all the social activities outside the lab, including the runs under the cold English rain!

Finally, I would like to say a huge thank you to my parents for their whole-hearted support and encouragement during all my academic studies. Thank you also to my family for their support and for trying to understand what I have been doing for the last 3 years. The final thanks go to Jean-Claude, Anne, Michel, and Evelyne for their encouragement and precious help for making the project of pursuing my studies abroad possible.

Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This study was financed with the aid of a studentship from the AHRC-funded 3D3 Centre for Doctoral Training (3D3 Research).

Word count of main body of thesis: 62191 words

Signed: .....

Date: .....

Abstract

Researchers have investigated harnessing computers as a tool to aid in the composition of music for over 70 years. In major part, such research has focused on creating algorithms to work with pitches and rhythm, which has resulted in a selection of sophisticated systems. Although the musical possibilities of these systems are vast, they are not directly considering another important characteristic of sound. Timbre can be defined as all the sound attributes, except pitch, loudness and duration, which allow us to distinguish and recognize that two sounds are dissimilar. This feature plays an essential role in combining instruments as it involves mixing instrumental properties to create unique textures conveying specific sonic qualities. Within this thesis, we explore harnessing techniques for the analysis and control of instrumental timbre and timbral combinations.

This thesis begins with investigating the link between musical timbre, auditory perception and psychoacoustics for sounds emerging from instrument mixtures. It resulted in choosing to use verbal descriptors of timbral qualities to represent auditory perception of instrument combination sounds. Therefore, this thesis reports on the developments of methods and tools designed to automatically retrieve and identify perceptual qualities of timbre within audio files, using specific musical acoustic features and artificial intelligence algorithms. Different perceptual experiments have been conducted to evaluate the correlation between selected acoustics cues and humans' perception. Results of these evaluations confirmed the potential and suitability of the presented approaches. Finally, these developments have helped to design a perceptually-orientated generative system harnessing aspects of artificial intelligence to combine sampled instrument notes.

The findings of this exploration demonstrate that an artificial intelligence approach can help to harness the perceptual aspect of instrumental timbre and timbral combinations. This investigation suggests that established methods of measuring timbral qualities, based on a diverse selection of sounds, also work for sounds created

by combining instrument notes. The development of tools designed to automatically retrieve and identify perceptual qualities of timbre also helped in designing a comparative scale that goes towards standardising metrics for comparing timbral attributes. Finally, this research demonstrates that perceptual characteristics of timbral qualities, using verbal descriptors as a representation, can be implemented in an intelligent computing system designed to combine sampled instrument notes conveying specific perceptual qualities.

Contents

List of Figures

List of Tables

List of Abbreviations

2D Two-Dimensional

ADAM Adaptive Moment Estimation

AI Artificial Intelligence

AIVA Artificial Intelligence Virtual Artist

ANN Artificial Neural Networks

ARTIST ARTificial Intelligence-aided Synthesis Tool

ATN Augmented Transition Network

Ato-ms Abstract Temporal Orchestration

CSAIL Computer Science and Artificial Intelligence Laboratory

CSL Computer Science Laboratories

CSS Concatenative Sound Synthesis

EMI Experiments in Musical Intelligence

FFT Fast Fourier Transform

FM Frequency Modulation

HNR Harmonics-to-Noise Ratio

Hz Hertz

ICCMR Interdisciplinary Centre for Computer Music Research

ILLIAC Illinois Automatic Computer

IRCAM Institut de Recherche et Coordination Acoustique/Musique

L-BFGS Limited-memory Broyden-Fletcher-Goldfarb-Shanno

Mac OS Macintosh Operating System

MDS Multidimensional Scaling

MFCC Mel-Frequency Cepstral Coefficients

MIDI Musical Instrument Digital Interface

MIR Music Information Retrieval

MIT Massachusetts Institute of Technology

MI Theory Theory of Multiple Intelligences

MPS Modulation Power Spectra

MRes Masters of Research

MS Millisecond

MUSICOMP MUsic Simulator Interpreter for COmpositional Procedures

MusicXML Music Extensible Markup Language

NSynth Neural Synthesizer

OQ Open Quotient

PCA Principal Component Analysis

PDF Portable Document Format

RBF Radial Basis Function

RBM Restricted Boltzmann Machine

ReLU Rectified Linear Unit

SACEM Société des Auteurs, Compositeurs et Éditeurs de Musique

SAD Sum of Absolute Difference

SNR Signal-to-Noise Ratio

SPORCH SPectral ORCHestration

SSD Sum of Squared Difference

SVD Singular-Value Decomposition

SVM Support Vector Machines

SVR Support Vector Regression

VAME Verbal Attribute Magnitude Estimation

WAVE Waveform Audio File Format

1

Introduction

1.1 Chapter Overview

This first chapter starts by introducing the different research areas along with the gaps and challenges that have motivated the investigations presented in this thesis. Then, the chapter details the research aims and questions addressed throughout the project, which defines the scope of this study. Next, the text presents an overview of the structure of the thesis. Finally, the last section lists the academic publications that have been produced as a result of the research presented within this thesis.

The structure of this chapter is as follows:

- 1.2 - Introduction
- 1.3 - Aims and Research Questions
- 1.4 - Thesis Structure
- 1.5 - List of Publications

1.2 Introduction

Musical timbre is a complex and multidimensional attribute of sound, whose standard definition is "*that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar*" [1]. This element of sound, defined as what it is not (loudness, pitch, and duration) rather than what it is, has been the subject of many discussions among the research community, see [2] and [3] for examples. Nevertheless, this sonic attribute, interdependent of pitch, loudness, and duration in the auditory sensation [4, 5, 6], is defined by acoustic and psychoacoustic properties [7, 8, 9]. A large amount of research on the subject has been produced since the first experiments conducted by Von Helmholtz at the end of the 19th century [10]. However, this element of sound still remains a notion that requires further investigation in order to completely define, understand, and also harness this complex characteristic into systems designed for musical purposes.

Timbre has an important role in tasks that combine several instruments. This attribute is of particular attention in orchestration, a compositional practice that involves writing for several instruments playing simultaneously. However, this research does not intend to address all the challenges of musical orchestration as it looks mainly at instrument combinations and it will not address the instrument technique constraints. Furthermore, this study has focused on processing audio recordings of instrument notes and orchestral pieces. Using these materials, many parameters related to orchestration are overlooked, such as instrument sound (e.g. body materials), performers (e.g. stylistic traits), and room acoustics for example. Nevertheless, the results of investigations into analysing and controlling instrumental timbre and timbral combinations would be part of the field of computer-aided orchestration, which looks at developing techniques and tools for aiding in aspects of musical orchestration. This field has burgeoned with the development of the technology and its increasing accessibility, where researchers and composers started to be more interested in using tools to aid harnessing sonic characteristics in orchestral compositions. The *Spectral Music* movement [11] initiated this compositional interest in the early 1970s, with composers such as Gérard Grisey and Tristan Murail in France, Johannes Fritsch and Clarence Barlow in Germany. The characteristic of this movement is the use of electronic devices, as well as computers, to analyse and represent timbral qualities of acoustic music.

Surprisingly, there are only a few computer-aided orchestration systems available. The majority of these have been developed in the last 15 years. This is perhaps due to the complexity of

musical orchestration and the limitations of the available technology at the time. However, the main breakthrough has been the result of research conducted at the Institut de Recherche et Coordination Acoustique/Musique (IRCAM), especially with their latest computer-aided orchestration system *Orchids*[1]. This tool proposes to output combinations of instruments that would match a target sound, using a pre-analysed and indexed database containing audio samples of orchestral instruments, offering an interesting approach to address some of the challenges of musical orchestration. Here, it aims to combine the data resulting from performing acoustical and psychoacoustical analyses on a target with the sound produced by combining orchestral instruments. While the solutions proposed by *Orchids* spectrally match the content of the target sound, the selection and combination of instrument notes can be sonically very different. When the user has a particular sound colour or a perceptual quality in mind, this outcome usually involves listening to several solutions, which can be a tedious and time-consuming task. One approach to overcome this issue is to propose a filter to narrow the solution space. This could be achieved by focusing on the sonic content of the generated instrument combinations, which would provide information for representing a desired perceptual quality. Therefore, the use of sonic characteristics representing perceptual properties could be an approach to allow for the manipulation and exploration of the different sonic textures created by combining instrumental timbres.

This study is looking at identifying techniques for the analysis and control of timbral properties resulting from combining orchestral instruments. The investigation will rely on processing of audio files in order to computationally represent perceptual properties and manipulate timbre characteristics. Such processes would certainly involve handling complex and varied types of data. The field of Artificial Intelligence (AI) has produced numerous efficient methods for solving a significant number of problems. Nowadays, AI is utilised to achieve various tasks and has found applications in almost every domain, with functions such as facial recognition [12], medical diagnosis [13], and speech-recognition systems with Apple's *Siri* for example. This small selection of applications illustrate the diverse and complex data that need to be processed by such AI systems. Following this observation, it would be interesting to investigate the potential use of different methods developed in the field of AI to harness some aspects of sonic perception, represented by musical timbre properties. Machine learning techniques, a subset of AI models, would provide an approach to analyse, identify, and predict the perceptual qualities of instrument timbre combinations from data resulting from the calculations of

[1] `http://forumnet.ircam.fr/product/orchids-en`

specific timbre properties. Such models could then be utilised to narrow the solutions output by some computer-aided orchestration systems. Furthermore, demonstrating the benefit of using AI methods would contribute in establishing a framework for the analysis and manipulation of the sonic qualities created by combining several instruments. Such methods related to the analysis of instrument timbre combinations could then be incorporated into computing systems designed to aid orchestral composition.

1.3 Aims and Research Questions

Thanks to the development of technology and its increasing accessibility, composers can now use sophisticated tools to aid in their metier. Such systems have produced excellent solutions to work with pitches, rhythm, and velocity, but harnessing musical timbre still requires further development. As briefly introduced in the previous section, timbre is a complex element of sound that is of particular interest in music, especially when combining several instrument sounds. Systems designed to address some challenges of orchestration propose to incorporate some aspects of timbre in their creative algorithms. However, due to the large number of potential combinations of instruments, such systems can output numerous solutions to a given target, which are not always matching the user's ideal type of sound.

The aim of this research is to investigate methods to harness aspects of music perception, represented by musical timbre properties, for the understanding and analysis of the sonic qualities produced by several instruments. Such methods could be used to address one aspect of orchestral composition related to instrumental timbre by establishing techniques for analysing, identifying, and controlling the timbre qualities resulting from sound produced by the combinations of instruments. This study focuses on audio sources using combinations of traditional Western instruments, however a similar approach could potentially be applied to any other types of instruments, either analogue or digital. The different research questions investigated within this study are introduced in the following section.

1.3.1 Research Questions

In order to achieve the aims of this study, there are different objectives that need to be filled. Therefore, the following research questions are addressed within this thesis:

- **RQ1**: Which sonic element can be used to evaluate and represent the perceptual quality of the sound emerging from a mixture of instruments by processing audio recordings of instrument combinations?

  One of the motives for composing orchestral pieces is the ability to create unique sounds from instruments playing simultaneously. Thus, composers may seek a specific type of sound or a texture, which is something being experienced from listening to instruments being played together. In the research initial stage, it is important to understand which

attribute to retrieve from an audio source to characterise the perception experienced from the listening process. The first research question also looks at finding a way to represent that perceptual quality in order to make it accessible to a broad audience. Research in the later stages of the investigation will draw on this observation to build methods for harnessing this aspect in computing systems designed to aid manipulating instrumental timbre combinations. This question will be addressed by evaluating and selecting approaches to retrieve perceptual quality from audio samples of combined instruments sounds and define a convenient representation for a wide audience.

- **RQ2**: How to compare timbral values resulting from the analysis of different acoustic properties?

  Several works have demonstrated the importance of acoustic and psychoacoustic properties in defining and retrieving timbre qualities. This involves performing various processes on an audio source. Data resulted from the various acoustic features have different meanings and different scales, which makes the values difficult to manipulate. Therefore, in order to be able to automatically classify an orchestral sound file according to its timbral quality, it is essential to design a comparative scale of the different timbral features that will be implemented in order to harness specific perceptual qualities in techniques developed for the analysis and control of instrumental timbres.

- **RQ3**: Which methods taken from the field of Artificial Intelligence (AI) could help in computationally analysing and identifying the perceptual properties of instrument combination sounds using values from musical timbre calculations?

  Considering timbre could help in retrieving perceptual qualities from audio files, and, added to a comparative scale, it would be possible to classify audio files of instrument combinations by processing their calculated timbre values. With no defined model to use such data, the use of AI frameworks to computationally represent timbre perception could allow for the design of an automatic classification method. This research question will be addressed by evaluating and selecting successful AI approaches using data collected from orchestral pieces in order to automatically classify audio samples of instrument combinations from their correspondent timbre values.

- **RQ4**: How to incorporate timbre properties into algorithms designed to generate combinations of sampled instrument notes?

An algorithm designed to generate sequences of instrument combinations can output a large quantity of solutions due to the combinatorial possibilities. To go beyond random processes, it is essential to add parameters and constraints to guide the combinatorial algorithms. After defining approaches to represent and automatically classify timbral properties, this study looks at harnessing these methods into algorithms designed to combine instruments. This last research question will be addressed by defining an approach to integrate timbre properties into a search algorithm, and therefore refining the search space by evaluating the sonic result of the instrument combinations. This approach will be illustrated with a computing system designed to generate combinations of string and brass notes using timbral descriptors.

1.3.2 Methods

This project uses an iterative approach to build an understanding on harnessing aspects of musical timbre into computing systems designed to analyse and control sonic qualities of instrument combinations by processing audio recordings of orchestral pieces and instrument notes. Firstly, various attributes were selected to build an initial system capable of computationally retrieving timbral content of orchestral audio files, including audio samples generated by the computer-aided orchestration system *Orchids*. Verbal descriptors of timbre qualities have been selected to represent the acoustic features, thus, alleviating the need of acoustics and psychoacoustics expertise to interpret and utilise the methods of timbre calculations. These methods were based on the literature that have defined the acoustic features that correlate with each attribute. This initial experimentation, combined with listener testing, informed the implementation of a method to automatically classify sounds emerging from mixtures of instruments according to their perceptual qualities, informed by their timbral content. Such findings have been integrated into a computing system designed to generate combinations of string and brass instruments matching specific timbral descriptors. The development of this system demonstrates the abilities of the methods proposed in this study, but also illustrate a potential application of harnessing analysis and control of instrumental timbre combinations for computer-aided orchestration systems.

1.4 Thesis Structure

This thesis is organised into eight chapters, including the present chapter. The rest of the thesis is structured as follows:

Chapter 2 introduces three important notions on which this research is built on. First, this chapter describes the human auditory and perception system, which are two key aspects in music listening. The second part introduces the concept of musical timbre by reviewing the significant works and applications produced by studies on this notion. The text also discusses the relation between audio perception and timbre. Finally, this chapter reviews the field of computer-aided orchestration, which has informed the initial developments presented in Chapter 4. It also highlights the different approaches that have been investigated in order to address some of the challenges for harnessing aspects of orchestration. This chapter intends to provide the acoustical and musical background knowledge and establish the current state of research in computer-aided orchestration in order to help the reader understand the presented study.

Chapter 3 presents different concepts of Artificial Intelligence (AI) that have been used in Music, and tries to suggest a definition in relation to the study presented in this thesis. This chapter starts with discussing and defining the notion of intelligence, and how it is computationally represented in AI. It briefly introduces the goals and applications of AI. Then, the chapter presents an overview of the use of AI in music, with some of the approaches having been used in a number of implementations presented in Chapters 5 and 6 of this thesis. The aim of this background chapter is to understand how AI could aid in answering the research questions mentioned in Section 1.3.

Chapter 4 describes the implementation of a computing system capable of automatically ranking orchestral audio files, representing instrument combinations, by performing specific signal processes to retrieve their timbre content. This chapter details the different verbal descriptors and acoustics features that have been used in the ranking system, along with technical details about its implementation. Then, the text details a perceptual experiment conducted in order to validate the chosen timbral approach. Results of the experiment are then analysed and discussed, along with the limitations of this system. Finally, this chapter discusses the potential applications and benefits of this initial timbral implementation, which supported further developments towards using these timbral attributes, particularly for the research detailed in Chapter 5.

Chapter 5 presents the development of an automatic timbre classification system for orchestral sounds representing instrumental timbre combinations. Built on the observations from the ranking system, described in Chapter 4, this chapter starts with discussing the challenges of developing a system able to automatically classify an audio file according to its perceptual quality. With no agreed metrics for timbre values of perceptual properties, the text details the creation of scales for the different timbral values, which was designed to overcome the issue of comparing different types of values. Then, the chapter continues with explaining the rationale for using different machine learning methods selected in the development of the automatic classification. The last part of this chapter discusses the different machine learning algorithms that have been implemented along with their performances for creating classification models.

Chapter 6 discusses the final implementation of this study, which is built on the observations from the research developments described in Chapters 4 and 5. It resulted in implementing a computing system capable of generating combinations of samples of recorded string and brass instruments presenting specific perceptual qualities. This chapter starts by defining the challenges to address for implementing a generative algorithm for instrument combinations. The text continues with a description of the sound database containing the string and brass instruments audio samples that are processed for timbre analysis and control of the creation of combinations. Here, AI techniques utilised for timbre prediction are detailed and discussed, which are then included in the search algorithm designed to generate combinations of instruments matching specific timbral descriptors that can be defined by the users.

Chapter 7 presents different sets of examples designed to illustrate the abilities of the generative instrument combinations system discussed in Chapter 6, which has utilised techniques identified from the research developments presented in Chapters 4 and 5. This chapter starts by defining the scope of the type of examples utilised to demonstrate the capabilities of the techniques developed throughout this study. Then, the text describes different examples of instrument combinations generated by the system, using different sets of instrument ensembles. These examples are then discussed in order to illustrate the abilities of the system for combining samples of recorded instruments' notes corresponding to desired perceptual qualities.

Chapter 8 concludes the thesis with a summary and discussions of the developments conducted in order to address the research questions listed in Section 1.3. This chapter also discusses the contribution to knowledge produced by the study presented in this thesis. Finally, the second part of the conclusion chapter lists different areas for further investigation in order to extend the advances provided by the research presented within this thesis.

1.5 List of Publications

Below is a list of the different publications that have been produced from the research presented within this thesis:

AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Towards intelligent orchestration systems**. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 671–681, Plymouth, UK, 2015 **[Reports on work from Chapter 2 and initial stage of Chapter 4, cited as [14]]**

AURÉLIEN ANTOINE, DUNCAN WILLIAMS, AND EDUARDO R. MIRANDA. **Towards a timbre classification system for musical excerpts**. In *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP)*, London, UK, 2016 **[Reports on work from Chapter 4, cited as [15]]**

AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Musical acoustics, timbre, and computer–aided orchestration challenges**. In *Proceedings of the 2017 International Symposium on Musical Acoustics (ISMA)*, pages 151–154, Montreal, Canada, 2017 **[Reports on work from Chapter 5, cited as [16]]**

AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **A perceptually orientated approach for automatic classification of timbre content of orchestral excerpts**. *The Journal of the Acoustical Society of America*, **141**(5):3723, 2017 **[Reports on work from Chapters 4, and 5, cited as [17]]**

AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Computer generated orchestration: Towards using musical timbre in composition**. In *Pre-Proceedings of the 9th European Music Analysis Conference (EuroMAC 9 – IX CEAM)*, Strasbourg, France, 2017 **[Reports on work from Chapters 5, and 6, cited as [18]]**

EDUARDO R. MIRANDA, AURÉLIEN ANTOINE, JEAN-MICHAEL CELERIER, AND MYRIAM DESAINTE-CATHERINE. **i-Berlioz: interactive computer-aided orchestration with temporal control**. In *Proceedings of the 5th International Conference of New Musical Concepts*

(ICNMC), pages 45–60, Treviso, Italy, 2018 **[Reports on work from Chapter 6, cited as [19]]**

AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Predicting timbral and perceptual characteristics of orchestral instrument combinations**. *The Journal of the Acoustical Society of America*, **143**(3):1747, 2018 **[Reports on work from Chapter 6, cited as [20]]**

2

Audio Perception, Timbre, and Computer-Aided Orchestration

2.1 Chapter Overview

The purpose of this chapter is to provide background information on key concepts related to the research presented in the thesis. It also aims to define the scope of this study and highlight the link between the three main notions that are introduced in this chapter: auditory perception, timbre, and computer-aided orchestration.

The first part of this chapter proposes a brief introduction to human audio listening and starts by describing the human auditory system in order to understand the way an audio signal reaches the brain and is subsequently interpreted. Then, an introduction to the general concepts of audio perception is proposed in order to further understand the interpretation of audio signals. This section ends with introducing the notion of musical perception, which is most relevant to this research project.

The chapter continues with discussions on the notion of musical timbre. Section 2.4 proposes an overview of the research on musical timbre, starting with its standard definition and is followed by discussing its different aspects in order to define this complex attribute. This section also focuses on the perceptual properties of timbre, which is the principal approach selected for the developments of this research.

Section 2.5 reviews the different developments produced by research in the field of computer-aided orchestration. The discussions of the different approaches that have been explored to address some of the orchestration challenges help determine potential improvements, which have

motivated some of the investigations presented within this thesis. The chapter then concludes with discussions about the different concepts and their relation to the presented study before finishing with a summary.

The structure of this chapter is as follows:

- 2.2 - Introduction
- 2.3 - Auditory System and Perception
- 2.4 - Timbre
- 2.5 - Computer-Aided Orchestration
- 2.6 - Chapter Discussions
- 2.7 - Chapter Summary

2.2 Introduction

When interacting with sound and music, hearing is the most active of the five traditional senses (namely sight, hearing, touch, smell, and taste) [21]. The ability to perceive sound from the vibrations produced by an object is the result of a complex mechanism, which is essential to experience various aspects of musical events. Our understanding of this sense has been the result of a long series of research on the subject, with important advancements occurring since the twentieth century. While this project does not attempt to model the human auditory system, and some of the presented concepts may be beyond the scope of this study, it is worthwhile to understand these hearing mechanisms when harnessing perceptual elements in a computing system designed to aid musical composition. Wiggins argues: "*the starting point for all music information retrieval (MIR) research needs to be perception and cognition, and particularly musical memory, for it is they that define Music*" [22]. Moreover, reviewing hearing processes highlights the relationship between music perception and the notion of timbre, which are important elements of this study.

2.3 Auditory System and Perception

2.3.1 Human Auditory System

The perception of sonic events starts with capturing sound, generated by mechanical vibrations, and is primarily achieved by the ears. This organ is the principal component of the human auditory system and has been the subject of many investigations for several centuries, such as research conducted by anatomist and physician of the sixteenth century Fallopio [23, 24], by Cotugno [25, 26], and also by Corti, known for describing the organ of Corti (named after him) [27, 28]. The significant amount of works has led to today's understanding of the physiology of the human auditory system.

There are three main components of the human ear: the outer ear, middle ear, and inner ear. The outer ear consists of the pinna, the visible part of the ear, and the external auditory canal. Their role is to collect and lead audio signals towards the middle ear. At the end of the auditory canal, the vibrations produced by the sound reach the tympanic membrane, or eardrum, then go through the malleus, incus, and stapes: three small bones located in the air-filled tympanic cavity. These components form the middle ear and function as a transformer matching the acoustical impedance between the air in the external auditory canal and the liquid in the inner ear. The last component of the human ear consists of the cochlea [29], which is filled with fluids and contains the organ of Corti [27]. It transduces the fluid movements into nerve signals via hair cells that go into the auditory nerve to be further processed by the brain. In summary, the ear performs an analysis of the audio signal before it is transduced and transmitted to the brain [30]. Figure 2.1 proposes a representation of the anatomy of the human ear.

Once the mechanical energy coming from the sound has been transduced by hair cells (acting as a mechanotransducer) into neural information, the message transits through the auditory nerve to the cochlear nucleus, located in the posterior part of the brain (called the brainstem). At this point, the information is in the central auditory nervous system. The signal then passes through several nuclei, via some located in the thalamus, in order to arrive in the primary auditory cortex (koniocortex) and the auditory association cortex (parakoniocortex), which are located in the temporal lobe of the cerebrum (see Figure 2.2). It is at this point that audio signal information is processed by the brain, which involves different cognitive processes. A detailed analysis and description of the auditory pathways can be found in [32]. Further details on the physiology of the global auditory system can be found in [33] and [30].

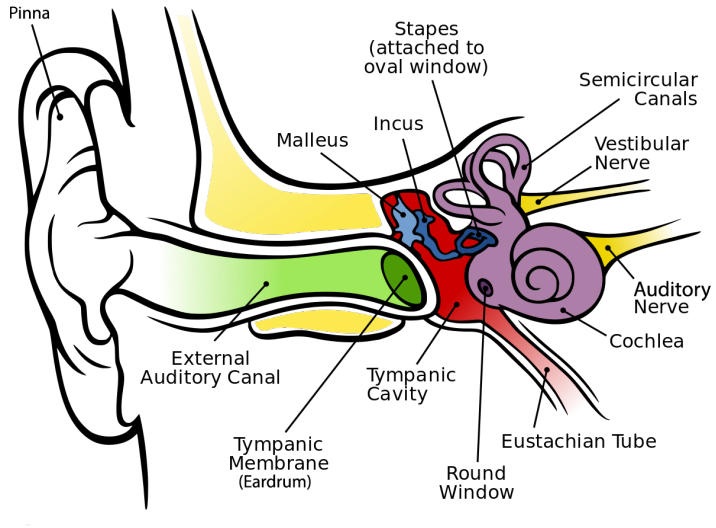

Figure 2.1: Anatomy of the human ear. The outer ear consists of the pinna and the external auditory canal. The middle ear is composed of the tympanic membrane (eardrum), malleus, incus, and stapes, which are small bones. The cochlea forms the inner ear, which transduces the signal into the auditory nerve to be conveyed to the brain for further processing. Adapted from [31] and published under the terms of the Creative Commons Attribution License ©2005 Chittka and Brockmann.

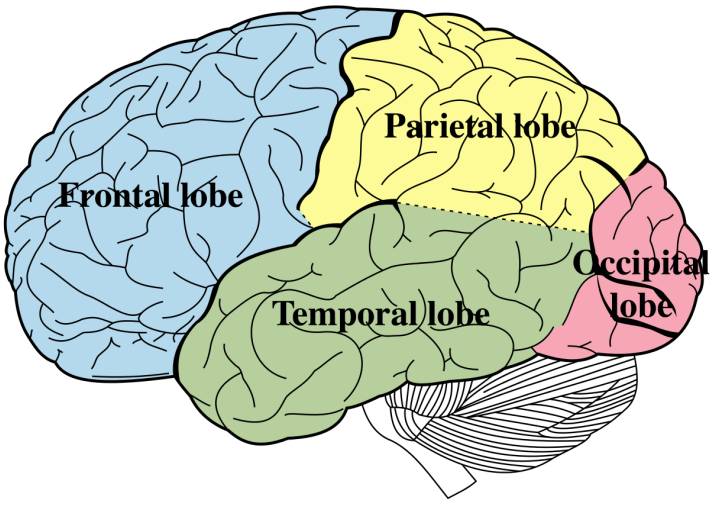

Figure 2.2: Representation of the four major lobes of the human brain. The auditory cortex is located in the temporal lobe, shown here in green. Image adapted from [34, Figure 728], under the terms of the Creative Commons CC0 licence.

2.3.2 Auditory Perception

Auditory perception is the ability of processing and interpreting audio information captured by the auditory system, and requires various cognitive processes and prior knowledge. First, it is essential to define the types of sound patterns that can be experienced by the human ears.

The simplest type of sound patterns are when properties of single sounds retain in a steady state. For example, in the case of sounds from musical notes produced by an instrument, only one property differs between the different notes—here, it is mainly represented by the fundamental frequency. Single sounds can also have changing states, meaning the acoustical properties vary over time, which makes the sound more complex in terms of information. For instance, using the sound produced by an instrument, in a crescendo or diminuendo, the intensity changes within the duration of the sound. Nevertheless, most of the sound patterns we experience are not solely composed of single sounds. Instead, they consist of a sequence of single-sound events that can differ in one or more dimensions. Sound sequences can vary in the number of elements and also in the order and timing of the elements [35]. For example, in Morse Code, the alphabet consists of a sequence of singe-sound elements, in which one dimension varies—in this case, the duration of each event. In a musical melody, the element of the sound sequence may vary in more than one dimension, such as frequency, intensity, and duration [36]. However, sound sequences, even with many variations among their elements, are not the most complex sound pattern. We usually receive several sources of sound sequences simultaneously. These sound sequences can contain correlated sound events or non-related events. For example, during a performance in a concert hall, we receive the sound of the instruments coming from the stage, but also the reverberation of that sound on the walls and ceiling of the venue. Per contra, conversing in a crowded room results in receiving sound sequences from speech and others from the background noise, which are independent sound information. Nonetheless, most of the time we are still able to separate the concurrent sound patterns and process the sound information.

Auditory perception can retrieve four main characteristics of sound: pitch, loudness, duration, and timbre [36]. Pitch allows us to scale a sound from low to high, and vice versa. It is related to the frequencies of the sound, obtained by measuring the oscillations of the waves, and expressed in Hertz (Hz). A sound is considered high when it contains high frequencies, and low when containing low frequencies [37]. Loudness is the intensity or volume, of a sound [38]. It refers to its physical magnitude and can be measured in sone (loudness N) and phon

(loudness level L) [39, 40]. Duration is the result of estimating the length of the sound event (difference between onset and offset) and is an attribute that applies not solely to auditory, but to all the human senses [41, 42]. Duration defines if the sound is short or long, and can be expressed in units of time (e.g. second (s), millisecond (ms)). The final characteristic is timbre, often referred to as the sound quality. This more complex characteristic (further explained in Section 2.4), allows us to distinguish and recognise voices, instruments, or sounds. For example, if the same pitch is played on a piano and a violin at the same velocity and duration, we can differentiate and recognise each instrument. While the other three characteristics are measurable, timbre is a multidimensional attribute and its estimation is more complex, as discussed in Section 2.4.

All the sound characteristics described above are essential in the auditory perception processes. First, the sound needs to be detected, which is possible if it has sufficient intensity (or loudness) to reach the ears, and if it is in the audible range (usually between 20 Hz to 20 kHz for humans [43, p. 163]). This first process allows the brain to detect the source of the sound and also if the source is moving, thus giving spatial information. The auditory perception is also able to discriminate the sound signals, which, for example, allows us to distinguish and hear the sound of a speaker from the background noise. Timbre is essential in identifying and recognising the source and type of a sound signal. For example, it allows us to identify if the sound is coming from an instrument, and, most of the time, which one. This attribute also contributes in determining if the sound is a voice and recognise if it is a known person. Finally, auditory perception, combined with other cognitive processes such as memory, allows us to understand the information of an audio signal. This process is important for communicating with others as well as understanding the meaning of a sound (a fire alarm, for example). All these auditory processes are also essential for experiencing music. Furthermore, it indicates that timbre is an important characteristic of auditory perception, which highlights the motives of harnessing this aspect into computing systems designed for musical creation.

2.3.3 Music Perception

In the previous section, different types of sound patterns that humans can experience have been described, and four characteristics of sound have been identified (pitch, loudness, duration, and timbre) as having an essential role in the different auditory perception processes (detection, discrimination, identification, and comprehension). While the perception of general sound events has been discussed previously, this section aims to highlight perceptual processes that

are more specifically related to music. This study does not intend to model musical perception, neither to detail and comprehend the different processes of understanding music. However, it focuses on the perception of one specific attribute of sound for musical purposes. Thus, a brief explanation of music perception may help the reader understand this study, and inform them of some decisions taken in the developments that will be detailed in Chapters 4, 5, and 6.

Music perception is highly linked to the field of psychoacoustics and the perception of the four attributes of sound detailed in the previous section. However, music perception encompasses a broad range of different fields to understand how we process musical information, such as neuroscience, cognitive science, psychology, and computer science. The definition of music is subject to many discussions, and there is no general agreement on what is music. Here, we will simply define music as a succession of sound events that has an artistic and emotional meaning. For instance, a fire alarm sound, which has a repetition of sound events, will not be considered as music on its own—even if this sequence of sonic events could potentially be incorporated into a musical composition. Also, this study focuses on music produced by traditional Western musical instruments, and, therefore, will not discuss synthesised sounds.

The perception of basic elements of sound, such as pitch, loudness, duration, and timbre, which are lower-level features, informs higher-level musical features such as melody, rhythm, and harmony. This involves an organisation and grouping of the sonic events. For example, a melody could be defined as the perception of a sequence of notes or pitches over a certain period of time. In this case, music perception is the result of an organisation of individual attributes over time. In orchestral pieces, several instruments can play simultaneously, however, these combined sonic events may form a single sound, even if the tones are the result of different instruments potentially playing different notes. Here, there is a grouping of the different pitches from various instruments, which results in forming a new and unique multi-instrumental note that is resulting in harmonic—frequencies that are at integer multiples of the fundamental frequency—or near-harmonic sound [44]. The closer the sound is to being harmonic, the stronger it will be perceived as a single note [45]. Similarly, the closer the fundamental frequencies of each sound are, the stronger they will be perceived as a single complex sound [46]. Time properties are also essential in the grouping of sonic events. Synchronous instrumental notes are more likely to be perceived as one than if one or more elements are asynchronous [47]. Sound and musical sequences can also be described using words and adjectives referring to ‘quality’, such as sharp, dull, noisy, or consonant. The perception of these sonic qualities involves different elements of sound and timbre is the one that plays an important role in distinguishing

these qualities [10, 48]. This notion, most relevant to the study presented in this thesis, will be further explained in Section 2.4.

Perceiving and understanding musical events are the results of several complex actions and information that are usually processed efficiently and quickly by the human brain. Furthermore, different parameters, such as the listener's cultural background and musical experience, to name but two, can have an influence in the perception and processing of music [49]. However, this research focuses mainly on the perception of sound qualities produced by notes from traditional Western instruments. Further information about music perception and the psychology of music can be found in [50], [51], and [52].

2.4 Timbre

In this study, the concept of musical timbre, and more specifically its perceptual properties, represents an important component of the research developments. This attribute is a complex property that has been studied for over a century and has produced a significant amount of work. However, some of its aspects are not relevant to the presented research. Therefore, this section proposes a review of selected timbral paradigms only. First, the standard definitions of this term are discussed. Then, different meanings of timbre are presented, with a review of the studies about different timbral paradigms, which informed the selection of the relevant timbre aspects that have been used in the developments presented in Chapters 4, 5, and 6.

2.4.1 Standard Definition

Different terms have been used to define the concept underlying musical timbre, such as *tone colour*, often used in treatises on orchestration [53, 54], or *sound colour* [55], which are translations of the German words *Tonfarbe* and *Klangfarbe*, respectively. However, in this thesis the word timbre will be utilised, which is the term generally accepted in the majority of literature on this subject.

Similarly to the other elements of the auditory system and aspects of sound presented in Section 2.3, timbre has been the subject of many studies since the nineteenth century. Von Helmholtz, a German scientist who is considered as one of the pioneering researchers in hearing science, mentions the concept of timbre in his influential book published in 1885 [10]. He states that

> "*when we hear notes of the same force and same pitch sounded successively on a piano-forte, a violin, clarinet, oboe, or trumpet, or by the human voice, the character of the musical tone of each of these instruments, notwithstanding the identify of force and pitch, is so different that by means of it we recognize with the greatest of ease which of these instruments was used*" [10, p. 19].

In a journal article published in 1934, Fletcher says that timbre is "*that characteristic of sensation which enables the listeners to recognize the kind of musical instrument producing the tone, that is, whether it is a cornet, a flute or a violin*" [56, p. 67], while Seashore defines timbre as "*that characteristic of a tone which depends upon its harmonic structure as modified by absolute pitch and total intensity*" [57, p. 97]. We can note that the definitions of timbre

proposed in these works have some variations. However, the standard definition of timbre, often cited in works related to this term, is "*that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar*" [1]. Notwithstanding, this definition can lead to different interpretations as it defines timbre by what it is not—not pitch, nor loudness—rather than what it is. Bregman suggests to rephrase the standard definition as "*we do not know how to define timbre, but it is not loudness and it is not pitch*" [58, p. 93]. In summary, timbre is a higher-order property emerging from loudness, pitch, and duration, which let a listener distinguish and recognise the sound from two different instruments. However, this definition suggests that unpitched sound, such as percussive instruments, would have no timbre [59]. Therefore, the definition proposed by Pratt and Doak as the sensation "*whereby a listener can judge that two sounds are dissimilar using other criteria than pitch, loudness, or duration*" seems more pertinent [60]. These standard definitions could suggest that timbre is simply the subtraction of pitch and loudness from the auditory sensation, and it assumes that timbre is entirely independent of pitch, loudness, or duration. However, several works have challenged this approach and suggested that these aspects of sound interfere in the auditory sensation [2, 4, 5, 6]. Thus, timbre is not just what is left over from pitch and loudness.

These imprecise definitions illustrate the complexity of this aspect of sound. Whereas loudness, pitch, and duration are clearly defined and can be scientifically quantified, measuring timbre represents another challenge. Furthermore, the notion of timbre is used to represent different properties, which makes its definition complicated. The following section tries to identify and illustrate the different meanings of timbre.

2.4.2 Types of Timbre

The previous section has illustrated the difficulties in proposing a unique definition of musical timbre. This is perhaps due to the different properties that are represented by the term timbre. This section aims to introduce the different meanings of musical timbre that are most relevant to the developments presented in this thesis.

Timbre as a musical element can be traced through the seventeenth and eighteenth centuries, where the 'colour' of the instruments started to gain interest. However, the first mention of timbre as a distinct musical element is perhaps in Berlioz's *Grand Traité d'Instrumentation et d'Orchestration Modernes* [53], a treatise on orchestration first published in 1844. Berlioz refers to orchestral composition as "*the use of [the] various sonorities and their application*

either to colour the melody, harmony or rhythm, or to create effects sui generis, with or without an expressive purpose and independent of any help from the other three great musical resources" [61, p. 6]. We can note that, here, timbre is defined as a musical colour. In another treatise on orchestration, published nearly fifty years after Berlioz's, Rimsky-Korsakov also mentions the importance of timbre: "*It is a great mistake to say: this composer scores well, or, that composition is well orchestrated, for orchestration is part of the very soul of the work. [...] One might as well say that a picture is well drawn in colours*" [62, p. 2]. Here, Rimsky-Korsakov suggests that orchestral compositions are not only notes on a musical score, but also the sound produced by the mixture of instruments that make up the orchestra. Therefore, timbre would refer to the texture of the sound and its perceptual characteristics. In a more recent treatise on orchestration, Piston emphases timbre as a musical element, and lists different techniques to fuse instrumental sounds together to create unique timbres [63, 64].

One meaning of timbre, related to the one usually mentioned in texts on orchestration, and perhaps the most frequently used to explain this notion, is the aspect of the sound that allows the listener to identify and recognise an instrument [60, 65]. For example, when different pitches are played with variable intensities on a violin, the listener will still be able to recognise the sound is from a violin. Here, there are variations in pitches, loudness, and duration but the timbre remains consistent [66, p. 411]. Nonetheless, we can also distinguish a middle C (C4) played on violin with a certain velocity and duration from a C4 played on piano in the same way. Here, pitch, loudness, and duration remain the same, but the timbre is different for each instrument. Timbre refers to the characteristic of the sound produced by a specific instrument [65, p. 238]. Note that sometimes, different instruments can be mistaken for others, due to high similarities in their sonic properties [10, p. 68]. The identification of a specific instrumental sound also depends on a previous learning process, in which individuals need training to classify and label instrumental sounds. Thus, a listener might not be able to identify instruments that have not been heard before. However, the listener will still be able to distinguish two or more different instrumental sounds.

Timbre can also refer to "*the way it sounds*" [67, p. 426]. Here, timbre is used to represent the perceptual characteristics, which can be related to the notion of sound's 'colour', as mentioned previously. Different words and adjectives, such as bright, dark, pure, or sharp, are often used to describe the 'quality' of a sound. These descriptors, which can be metaphors and analogies associated with the senses of vision and touch, are in fact terms that encompass multiple sonic attributes, mainly consisting of acoustical properties. This notion of timbre is

an important aspect of the research developments presented in this thesis. Studies about the relation between verbal descriptors and quantifiable acoustical attributes, mainly the result of psychoacoustical research, will be further explained in the next section.

The two timbral paradigms introduced in this section are of particular interests in harnessing timbre properties for instrument combinations. Firstly, timbre represents the distinct characteristics of the instruments, referring to the concept of *source timbre* [68]. Here, each instrument has its own timbral characteristics, referring to the typical sound it produces. Secondly, the possibility of combining several instruments offers the ability to create unique sound textures. Here, timbre encompasses the different perceptual properties that allow the listener to experience these textures emerging from the instrumental mixtures. The combinations of the instrumental timbre characteristics allow for the creations of unique sounds that contribute in experiencing specific perceptual effects.

2.4.3 Psychophysical Studies of Timbre

In Section 2.3, it was mentioned that pitch and loudness can be scientifically quantified, with frequency and with the amplitude, respectively. In regards to timbre, the measure of its sensation is more complex. As it has been mentioned previously, timbre encompasses multiple sonic attributes. This section aims to review the considerable amount of research that has been produced in order to understand what elements of the sound relates to the sensation of timbre. It also presents the different approaches utilised to develop our understanding of this complex and multidimensional component of sound.

A note added to the definition proposed by the American Standards Association attempts to extend the explanations of timbre by specifying that "*timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus*" [69]. In [7], Schouten states that there are at least five important acoustic parameters to determine timbre:

1. The range between tonal and noiselike character

2. The spectral envelope

3. The time envelope in terms of rise, duration, and decay (ADSR - attack, decay, sustain, release)

4. The changes both of spectral envelope (formant-glide) and fundamental frequency (micro-intonation)

5. The prefix, or onset of a sound, quite dissimilar to the ensuing lasting vibration

We can see that with these two statements, timbre encompasses many different properties. However, this information can be divided into two categories: resulting from the frequency spectrum (spectral information) and from the temporal characteristic of the sound (temporal information).

The relation between the timbre of instrumental tones and their frequency spectrum was already suggested by Ohm in 1843 [29, 70]. However, this assumption, known as Ohm's acoustic law, has been further investigated and elaborated by von Helmholtz in 1885 [10]. Ohm's acoustical law suggests that the sensation of timbre in musical sounds is the result of the analyses of the individual tones that compose a complex sound. Thus, "*for any complex wave-form the ear 'hears out' the simple harmonic components, the same components that Fourier's analysis or resonance would give*" [70, p. 326–327]. Both Ohm and von Helmholtz stated that the sensation of timbre was solely dependant on the patterns of the harmonics and fundamental frequency creating the individual tones, excluding the difference in phase. This assumption has been questioned since and it has been demonstrated that changes in phase of the component's harmonics can be perceived by the listener [71, 72] with, however, relatively minor effect.

Several works have demonstrated the connection between timbre and spectral information, following the first experiments by Ohm and von Helmholtz. The boosting or attenuation of the amplitudes of partials will result in changes in timbre [73, 74, 75]. For instance, Risset and Wessel have shown that a change in the spectrum around 2000 Hz modifies the 'presence' of a sound [76]. In fact, even the modification of a sine wave's frequency will result in changes in its timbre [77, 78]. The link between timbre and spectrum is also highlighted with formants. Here, a formant refers to a peak in the amplitude of certain frequencies, which is often associated with a sound's resonance [76]. In music, a formant can be defined as a resonance of a note's harmonic [79]. A change of pitch from an instrument will result in a difference in the fundamental frequency of the sound and produce a different spectral envelope. However, the formant frequencies will remain similar, which suggests that it is the characteristic of an instrument. This would allow the listener to identify the origin of the sound—recognise the instrument [55, 73]. Furthermore, it has been demonstrated that invariances in the structure of the formants lead to perceiving timbral similarities rather than invariances in the spectral envelope [80, 81]. Timbre is therefore linked with the spectral energy distribution of the sound.

Although it has been established that perception of timbre is linked with properties of the frequency spectrum, which results in the steady state of the sound as suggested by works fol-

lowing von Helmholtz's publication from 1885, the sensation of timbre also emanates from temporal characteristics of the sound. Musical sounds mostly evolve in time and the spectrotemporal variations of the sound provides important information about timbre characteristics [73, 74, 82]. For instance, the onset and offset of a sound provide important cues for identifying its source. This has been demonstrated in an experiment by Stump, where removing the attack of a note proved to significantly reduce the ability to identify the instrument [78]. The relevance of onset and offset information in identifying instruments has been further investigated in [83]. Furthermore, the amplitude envelope also contributes in identifying the source of the sound. For instance, when a note is played backward, its amplitude envelope is reversed, which causes difficulty in recognising the instrument, as demonstrated by George in [84].

The role of sound onset, offset, and amplitude envelope has been further investigated by Berger [85] using recorded tones of wind and brass instruments. In this experiment, listeners were presented different recorded versions of instrumental tones: unedited recordings, onsets and offsets removed, reversed recordings, and low pass filtered recordings. This study suggested that, unsurprisingly, unedited tones produced the best score, while the low pass filtered stimuli appeared to be the most difficult recordings to identify. Also, this study suggested that onsets and offset characteristics tend to have a more important role in identifying instruments than the amplitude envelope. These results have been confirmed in a study conducted by Saldanha and Corso [83], which used string instruments in addition to wind and brass instruments. With these studies, we can note that spectrotemporal variations of the sound also inform on the sensation of timbre.

The perception of timbre, unlike the sensation of pitch and loudness, is multidimensional. A vast amount of research has been focusing on establishing the number of timbre dimensions and identifying the acoustical characteristics of the signal in correlation with the timbre spaces. These perceptual researches mainly consisted of rating the similarity (or dissimilarity) between sound pairs and in identifying instruments. Most of the studies about sound similarity used stimuli of instrumental sounds, either natural [86, 87, 88, 89] or synthetic [90, 91, 92], which were judged by human subjects. The ratings of perceptual timbre similarities were analysed using the technique of Multidimensional Scaling (MDS) [93, 94, 95] in order to produce a geometrical representation of the timbre spaces, in which points are representing the different stimuli, and the distances between the points refer to their dissimilarities [67, 96, 97, 98]. In [99], the authors have investigated the dissimilarities of 42 timbres and obtained a five-dimensional MDS space. These studies on timbre spaces have demonstrated the link between

some acoustic features and timbre qualities. The spectral centroid, which corresponds to the mean of the spectral energy distribution in the spectrum, has usually been consensually correlated to the perception of *brightness* [100, 101, 102]. Research has suggested other attributes, which includes spectrotemporal variations [87, 90], attack and decay transients [75], spectral shape [92], and also spectral irregularity [91]. We can see that acoustical features representing timbre spaces vary among these works. Thus, it has been suggested that timbre spaces might depend on the context and the choice of the stimuli [97].

In the previous section, we have seen that timbre can be used to represent the perceptual characteristics of a sound or the 'colour' / 'quality' of a sound. Several studies have investigated the use of verbal descriptors of timbre qualities (or timbre semantics), with the objectives of finding their acoustic correlates. One of the first studies on using adjectives to describe timbre was conducted by Lichte [103]. However, one of the most complete early studies has been conducted by von Bismarck [48, 104], in which he suggested that four scales (dull-sharp, compact-scattered, full-empty, and colorless-colorful) can be used to define the timbre of individual instrument tones. Following von Bismarck's experiments, timbre semantics have been the focus of many studies [60, 105, 106, 107, 108, 109]. Such experiments have usually used Verbal Attribute Magnitude Estimation (VAME) [110] or semantic differentials [111] ratings. The VAME rating scale quantifies the applicability of the descriptor on the sound (e.g., warm $\leftrightarrow$ not warm), whereas the semantic differentials refers to using a scale of opposite descriptors (e.g., full $\leftrightarrow$ empty). The use of semantic differentials scales has been questioned by some researchers stating that the selected descriptors may not necessarily be opposite of each other [107, 110]. Nevertheless, these two rating scales, used in perceptual experiments, have proved to contribute in understanding timbre semantics [60]. The verbal descriptor of timbre qualities that has been consensually agreed is *brightness*. Von Helmholtz had already suggested the term 'bright' or 'brilliant', in his study published in 1885 [10]. Perceptual experiments, using either VAME [107, 108, 112] or semantic differentials [113, 114], have demonstrated the applicability of using *brightness*. Different studies have shown that the acoustic correlate of the descriptor *brightness* is the spectral centroid of the sound [87, 102, 108, 115, 116]. An extensive list of verbal descriptors and their correspondent acoustic correlates can be found in [117, Section 2.2]. Furthermore, the verbal descriptors that are directly related to the developments presented in this thesis will be further detailed in Chapter 4.

The studies on timbre have also focused on identifying the perceptual characteristics for musical instrument recognition, which is another meaning of timbre—the specific sonic char-

acteristic of an instrument. Several studies have shown that the sensation of instrumental notes is the result of different sound characteristics. As it has been mentioned previously, some of the timbral information is derived from the frequency spectrum of the sound. Here, information related with formants seems to have an important role in the perception of instrumental tones [118, 119]. Characteristics related to time variations, such as attack and decay transients [75] and spectral flux [92], are likewise important in the sensation of specific instrumental tones. In his study [120], Kendall suggested that musical context had an important role in a human's ability to recognise musical instruments. Here, Kendall compared the identification performances between single instrumental notes and whole-phrase, the latter producing better identification scores. This study identified the role of the information over time in the perceptual mechanisms and not only the spectral information of individual note. This role was confirmed in a study conducted by Sandell [121], in which the subjects were presented recordings of arpeggios with variable number of notes. Unsurprisingly, the instrument identification results were higher when more notes were presented.

The investigations on timbre characteristics of musical instruments have been the subject of many works in relation to human sound discrimination abilities [122, 123]. Moreover, the development of the computer has lead in developing systems for automatic instrument recognition. Some of the early systems used spectral information [124] or temporal information [125], however, a combination of both types has been the standard in most systems [126, 127]. These studies have suggested that Mel-Frequency Cepstral Coefficients (MFCC), first used in speech recognition [128], tend to be the most important features [129]. Brown [124] demonstrated his system, which used only MFCCs, on a set of oboe and saxophone sounds, and obtained near-perfect recognition. The contribution of MFCCs in instrument recognition was confirmed in a study conducted by Eronen, which used a set of sounds from 30 orchestral instruments [130]. More recently, it has been proposed that Modulation Power Spectra (MPS) could have a prominent role in the perception and classification of musical instrument sounds [99, 131, 132, 133]. Here, MPS corresponds to a two-dimensional (2D) Fourier transform of the time-frequency representation of a sound signal (spectrogram) and represents the temporal and spectral periodicities of the signal [99, 134]. Recent studies using the MPS approach suggest that specific spectrotemporal modulations are characteristic of instrument timbres, which would help automatic source recognition tasks [135, 136]. However, the work presented in this thesis does not intend to identify musical instruments automatically, interested readers can find a comprehensive overview in [137, Section 3.5].

Musical timbre has been the subject of many studies since the first experiments conducted by von Helmholtz in the second half of the nineteenth centuries. The research on this sonic attribute has produced significant understandings of this complex concept and helped in defining the meanings and properties of musical timbre. However, the majority of these studies have used sounds from clean recordings of isolated individual notes, whom experimental conditions may not reflect a realistic context. Indeed, in 'real-world' conditions, sounds are composed of multiple sonic events that are perceived in parallel. Similarly, in music, listening to an individual note in poly-instrumental pieces may be difficult and the characteristics identified for monophonic timbre may not apply in this context. This aspect is also important in the presence of instrument combinations, which is most relevant to this study.

2.4.4 Polyphonic Timbre

Many studies have identified acoustics cues for the perception of timbre. However, most of these studies have used sound samples of isolated individual notes as stimuli. In realistic contexts, and especially in musical contexts, sounds are composed of different layers of sonic events (in regards to this research multiple instruments), which has been named as polyphonic timbres [138, 139]. Gjerdingen and Perrott suggest that polyphonic timbre is a collection of variations of the spectral and temporal information from the audio signal [140]. This different sonic information may influence the perception of timbral quality either for the task of identifying instruments or in perceiving and describing the 'colour' of the sound.

While the perception of monophonic timbre has produced a considerable amount of works, polyphonic timbre has been the focus of less research. One of the first study focusing on the timbre perception of sounds from multiple instruments has been conducted by Kendall and Carterette [89]. In this study, the authors have compared the results of the perception of individual notes using dyads (two notes played simultaneously). The dyads were composed of single tones or simple melodies played in harmony and synchronously. Their results suggest that the perception of timbre combinations could be modelled as a linear summation of the individual timbres, which would result from the individual sound attributes. However, many timbre characteristics are the result of spectrotemporal variations, which are not the result of linear functions. Thus, the usual acoustic features suggested by the studies on monophonic timbre could not be applied to signals of polyphonic timbres, and therefore would require source separation, which is not necessarily a simple process [141, 142]. This study also suggest

that the perception of the timbre of dyads sounds varies according to the instrument pairs and that the musical context also provides essential information for identification.

Sandell conducted different experiments related to the perception of combined instrumental sounds [143, 144]. Similarly to Kendall and Carterette's experiment, Sandell investigated instrumental dyads, played in unison at interval of a minor third, and the role of the spectral centroid in perceiving the overall timbre as *bright* or *dark*, as demonstrated in monophonic timbres studies. Results of these experiments suggest that spectral centroid of the global sound, that is resulting from the overall spectral envelope, may not be enough to capture the timbral quality (in this case the attribute *bright/dark*) of the sound creating the instrumental tone combinations. Tardieu and McAdams also investigated the perception of dyads, with a particular focus on impulsive and sustained tones [145]. Their first experiment sought to identify the characteristic that create the blendedness of dyads, while the second experiment focused on identifying the factors in perceptual dissimilarity ratings between dyads. This study confirmed, as mentioned previously, that time-varying properties have an important role in timbre perception [92, 144] and auditory sensation [47, 146]. Similarly, spectral information, such as spectral centroid, spectral spread, and spectral envelope has a significant function in the perception instrumental tones combinations. Finally, this study has also demonstrated that dyads blending is primarily influenced by impulsive instrumental sounds, whereas the sustained instrumental tones influence the overall timbre.

Aucouturier has conducted different experiments in order to model polyphonic timbre [139]. These studies have investigated the measurement of timbre similarity between polyphonic musical sounds, which suggested that timbral information retrieved from the overall sound may not provide efficient data. Therefore, it would require a source separation process in order to retrieve accurate timbre cues, as suggested in Kendall and Carterette's study [89]. Furthermore, these experiments have suggested that the perception of polyphonic timbre is the result of high-level cognitive processes, which would also integrate notions from prior knowledge, cultural expectations, and musical context, instead of an immediate sound perception process as suggested in monophonic timbre studies. Such conclusions imply that polyphonic timbre is exponentially more complex than monophonic timbre and timbral characteristics identified in previous works may not apply in this situation. Thus, this study suggests that research on polyphonic timbre will require investigating different fields, other than only psychoacoustics, in order to represent this notion.

In [147], Alluri and Toiviainen have explored the verbal and acoustic correlates of polyphonic timbre. This study consisted of two experiments. The first investigation attempted to identify verbal descriptors that could be applied to polyphonic timbre, following similar semantics used in work on monophonic timbres, as mentioned in the previous section. For the second part of their study, they selected verbal descriptors of timbre spaces identified in the initial experiment and investigated their potential acoustic correlates. The results of this study suggest that the perceptual information identified in research on monophonic timbres may be applied to compute perceptual characteristics of polyphonic timbres. However, in the case of perceiving *brightness*, which has been consensually identified as correlated to the spectral centroid, the results of the study showed that spectral centroid did not highly correlate with the sensation of *brightness*. The reason for the different correlation could be related to the cognitive listening process, as suggested by Aucouturier [139]. Different musical information may influence on the sensation of *brightness*, such as pitch content as suggested by Alluri and Toiviainen's conclusions [147]. This study also suggests that the length of the stimuli may influence the sensation of polyphonic timbre, which could be related to the importance of musical context in perception of the instrumental timbres [120]. Results of this investigation demonstrated that MFCC, successfully used in instruments recognition [137, Section 3.5], may not have an important role in polyphonic timbre spaces. Finally, this study also suggests that spectrotemporal variations may provide significant information in the perception of polyphonic timbres. An extension of these studies on polyphonic timbre conducted by Alluri can be found in [148].

Unlike monophonic timbre, as highlighted in the previous section, polyphonic timbre has received less attention in perceptual experiments. The results of the studies on polyphonic timbre have suggested mixed conclusions, which do not provide a clear understanding and framework. These mixed results could be explained by the use of different types of stimuli, which, as mentioned in monophonic timbre studies [97], may influence the perception of different timbres. Another explanation could be that perception of global sounds would require discrimination processes (or source separation) combined with high-level cognitive actions, as suggested by Aucouturier [139]. Nevertheless, further investigations on polyphonic timbre are required in order to confirm or reject the assumptions suggested by the few available studies. Such advancements would benefit in methods related to the perception of instrumental sound combinations in which the notion of polyphonic timbre is an important component. Each instrument has its own timbral characteristics. Therefore, combining instruments will result in

creating polyphonic timbres, demonstrating the importance of this notion for the research developments presented in this thesis.

2.5 Computer-Aided Orchestration

This study aims to establish techniques for the analysis and control of instrumental timbre and timbral combinations. While this approach is not addressing many aspects of musical orchestration, such as instrument technique constraints, it is based on works from the field of computer-aided orchestration and the results of this investigation would benefit this research area. Therefore, this section presents a review of the different approaches that have been investigated for harnessing aspects of musical orchestration in computing systems.

The first computer-assisted orchestration systems can be associated with the *Spectral Music* movement, a term coined by Dufourt in [11], initiated in the 1970s. Composers such as Gérard Grisey and Tristant Murail, to name but two, initiated this movement in France in the early 1970s, with the ensemble *L'Itinéraire*, based at the *Institut de Recherche et de Coordination Acoustique/Musique* (IRCAM) in Paris. During the same period, some composers such as Johannes Fritsch and Clarence Barlow, from the Feedback Studio situated in Cologne, were also part of this new musical movement. Composers in this movement were focusing on the timbre quality of acoustic instruments and synthesised sounds and started to use electronic devices and computers to assist them in different aspects. One of the first pieces associated with spectral music is entitled *Partiels*, which was composed by Grisey in 1975 for 18 instruments. Here, a spectral analysis was realised on the trombone using an electronic sonogram. Following this experiment and the development of the technological and scientific knowledge and tools, various composers started to use computers to help them compose music for ensembles.

In his piece *L'Esprit des dunes* (1994) realised at IRCAM, Murail started his composition by analysing fragments from different sources such as diphonic Mongolian singing and Tibetan singing. The material of this composition, for an ensemble of 11 instruments and electronics, was generated by a spectral analysis of the aforementioned sources. He used an analysis program developed for additive synthesis then constructed a database of these analyses to be evaluated and modified with libraries he developed in the visual programming environment *PatchWork* [149]. Some composers used speech analysis for their compositions. Clarence Barlow developed a technique called *Synthrumentation*, which consisted of doing spectral analysis of speech and then mapping these analyses to acoustic instruments [150, 151]. Malherbe also used voice analysis techniques for some of his compositions. In *Locus* (1997) [152], a piece for four voices and electronics, Malherbe recorded singers using two microphones placed at two

different distances in order to have two different characteristic recordings. After the segmentation, smoothing, and normalisation of the recordings, Malherbe applied a spectrum analysis to obtain a representation of the material in the form of a sonogram. Then, a detection of partials was applied and the most prominent ones were selected. These data were subsequently input into *PatchWork* and transcribed into symbolic representations for ease of manipulation. For the rhythmic representation and manipulation, Malherbe used *Kant*—a rhythm quantification program developed at IRCAM by the Computer Assisted Composition research group[1] [153].

For his musical piece entitled *Metal Extensions* (2001), Maresz used a set of techniques combining handwriting and computational processes. In [154], he described his process as follows:

> *"Selection of the region of sound to orchestrate from the electronic sound file, placement by hand of markers on the region within the sound file that interested me, for a chord-sequence analysis with AudioSculpt (peaks), inharmonic partial analysis on the totality of the sound file in the same programme, transcription of the given results into symbolic notation in OpenMusic and finally, realization of the final score by hand."*

As a summary, several composers began to see the potential of the computer to help them orchestrate musical ideas or at least using computers in some of the different compositional processes. As seen with the spectral movement, computers were used for frequency analysis and representations of audio signals for examples. Composers used software, such as *AudioSculpt*, *PatchWork* or *Macaque* [155] to name but three, to analyse sound and for the representation and manipulation of the symbolic view of orchestration. With technological advances and experimentations from several composers, the idea of developing computer systems for addressing different aspects of orchestration started to arise in research groups.

Surprisingly, there are only a few computer-aided orchestration systems available. The majority of these have been developed in the last decade. This could be due to the complexity of orchestration and the limits of the available technology at the time. One of the first attempts was a tool developed by Rose and Hetrik [156]. They proposed a system that analyses a given target sound and it outputs an orchestration that tries to approach the target file. Their algorithm used a Singular-Value Decomposition (SVD) method for the proposition of new orchestrations using the spectrum of the target sound. The SVD approach, which involves performing factorisation

[1] `http://www.ircam.fr/repmus.html`

of complex data matrix, is interesting in terms of low calculation costs, and the solution is the nearest to the target sound. However, this approach does not take into account the position of the orchestra and the problem of instruments combinations.

Psenicka proposes another approach, with his program called *SPORCH* (short for SPectral ORCHestration) [157]. Like the system proposed by Rose and Hetrik, this program analyses a target file and outputs the orchestration solutions in the form of a list of data. This data comprises instrument names, pitches, and dynamic levels in order to create timbre and quality that fit the target file. Psenicka decided to perform the searching algorithm focussing on instruments, instead of doing it on sounds. Hence, the system is divided into two parts: the instruments database and the orchestration function. The database first needs to be built in order to run the program, which contains a list of instruments with pitch range, dynamic level range, and the most significant partials associated with the instruments (see [157] for more details). To find the orchestration solutions, Psenicka uses an iterative matching algorithm to establish the combination of instruments that fit the original file. The algorithm extracts the peaks of the target sound, and then compares them with each instrument in the database in order to select those closest in frequency. This approach has a low computational cost, and the instrumental composition of the orchestra is incorporated in the matching algorithm. However, this method tends to output simple orchestrations and only the solution that best matches the target file. Therefore, it discards all other solutions that could be more interesting in terms of musical ideas.

Another attempt is the system developed by Hummel [158]. Like Psenicka's method, he used an iterative algorithm, but instead of analysing spectral peaks, the program works with spectral envelopes—the frequency-amplitude derived from a Fast Fourier Transform (FFT) analysis. As with the two previous systems, this system analyses a target sound, retrieving its spectral envelope and then searches iteratively for the best approximation. Hummel says his system works better with non-pitched sounds (e.g., whispered vowels). This is due to it using spectral envelopes instead of spectral peaks. Hence, the perception of the pitches resulting from the solutions can be different from the pitches of the target file.

IRCAM addressed the question of computer-assisted orchestration in 2003, initiated by a research project proposed by Maresz [154], whose works have been mentioned earlier. Their first computer-aided orchestration system, named *Orchidée*, was the result of two Ph.D. studies conducted by Tardieu [159] and Carpentier [160]. They respectively addressed the problem of instrumental sound and timbre content analysis, and the rapid increase of the possible solutions

that can be produced by the matching algorithm. Their system extracts different audio information from a target sound and from the audio samples contained in the instrumental notes database. This information is used as data for the combinatorial algorithm developed to match the target sounds [161, 162]. In this computer-aided orchestration system, the algorithm does not output only the best solution but rather a selection of optimal solutions, which is an advantage as it can propose different orchestrations for one target sound. However, this version did not consider the temporal problems of orchestration. The system proposes only static orchestration solutions and works best with static and harmonic target sounds. The composer Jonathan Harvey, assisted by researchers and computer music designers from IRCAM, was one of the first composers to benefit from this new computer-assisted orchestration system. He used it for his composition *Speakings* (2008), which is for live electronics and a large orchestra [163]. Here, Harvey recorded three vowels, as mantra sung: *Oh/Ah/Hum*. Then, his idea was to input these recordings into *Orchidée* in order to try to imitate the sound produced by the sung mantra and obtain an orchestration for an ensemble of 13 instruments.

IRCAM's system *Orchidée* evolved into a new version, under the name *Ato-ms* (short for Abstract Temporal Orchestration), which was the result of a third Ph.D. study on the subject conducted by Esling [164]. One of the major improvements of this version was the management of time. Here, the system was designed to generate orchestration solutions within a time space as opposed to static. Another improvement was the use of a multi-objective and time-series matching algorithm, which creates a dynamic time warping algorithm. Furthermore, in this version, the user was able to design envelopes for the audio features, thereby able to create abstract targets. However, according to Maresz [154], the solutions "*suffer from a lack of quality in their timbral accuracy*", and the two versions of their computer-aided orchestration system only address the problem of timbre matching.

In November 2014, IRCAM released a completely new version of this system, named *Orchids*[1]. This standalone application implements the best features from its predecessors and integrates new improvements. It proposes abstract and temporal orchestration and is also optimised for timbral mixture. Like the aforementioned systems, the user inputs a target sound, but in *Orchids*, they also can design a complete abstract target by selecting and shaping various psychoacoustic descriptors. Some of the acoustic and psychoacoustic descriptors are based on works from the project *CUIDADO* [165], also conducted at IRCAM. *Orchids* also includes a database of over 30 orchestral instruments whose audio samples are pre-analysed and indexed

[1] Orchids is available at `http://www.forumnet.ircam.fr/product/orchids-en`

by the program. It is also possible to extend the sound database by simply adding a folder into the program, which will then be analysed and indexed. The user also can define the instruments they want to include in the orchestra. Moreover, as *Orchids* integrates the notion of spatialisation of the orchestral space, the user can position the instruments. The system first analyses the defined psychoacoustic features from the target. Different matching algorithms are available that are specific to the type of solutions the user wants or the most appropriate for the type of audio target (see [166] for more details). *Orchids* usually proposes several orchestration solutions in the form of a musical score, using the *Bach* library[1]. The user is also able to listen to the solutions before exporting the ideal orchestrations. This is possible due to the audio samples contained in the database. Furthermore, the user has the ability to start constructing and editing the composition directly inside the program using the score editing options. Then, once the user finds an orchestration that corresponds to their ideas, it is possible to export the results within different file formats, namely audio file, *bach* file, MIDI (short for Musical Instrument Digital Interface) file, *OpenMusic* file, and *PWGL* file.

IRCAM's latest program *Orchids* presents promising improvements for solving the challenges of computer-aided orchestration and it is set to be a powerful tool. However, one potential problem with *Orchids*, and in some other systems mentioned previously, is the numerous solutions they produce to orchestrate a given sound, which can be beneficial for inspiration, but at the same time it is tedious and time-consuming to go through all the solutions. An important step in computer-aided orchestration could be to focus on how to personalise systems to a user's style in order to offer more accurate orchestration solutions. The ability to have all possible solutions should be kept because chance or surprise is part of the compositional process, but computer-aided orchestration systems could be more efficient in regards to achieving composer specific musical ideas. Therefore, including an option to filter the numerous solutions output by such systems would offer a way to overcome the large instrumental combination possibilities.

The use of words to describe musical characteristics is a common practice in many aspects, whether for composition, music production tasks [167, 168] or simply by listeners. Many investigations on semantic and music (or more broadly audio) have been conducted, which led in the development of the field of semantic audio. A comprehensive review of these studies is beyond the scope of this thesis. However, further details about applications of semantic audio can be found in [169]. Nevertheless, the use of audio descriptors has also recently emerged

[1] `http://www.bachproject.net`

in some computer-aided orchestration approaches, such as [170, 171] for example, which followed ideas suggested in [14]. This indicates that the use of verbal descriptors could be utilised in guiding search algorithms designed to generate instrumental combinations towards audio qualities that can not be covered by spectrum matching algorithms, and therefore offer a way to filter the potentially large number of combinations by presenting only those having specific perceptual qualities. This could also go toward harnessing the creation of sound textures, an important aspect of timbral combinations.

The most recent attempt at developing a computer-aided orchestration system is perhaps *Live Orchestral Piano* [172]. Here, the system is built upon an orchestral writing technique proposed by Piston, which consists of first writing a harmonic and rhythmic structure for piano and then adding the other voices for all the instruments that compose the ensemble [63]. This technique is referred as projective orchestration [164]. The authors observed that this process is not solely a distribution of the piano notes across the other instruments, but rather involves a manipulation of timbres based on the existing harmonic and rhythmic structure from the piano score. Timbre as a structuring approach has been proposed by McAdams [173]. Their generation of projective orchestrations algorithm is based on a probability distribution model. Here, Crestel and Esling analysed a repertoire of different projective orchestrations, including works by Liszt who reduced Beethoven symphonies to piano for example, in order to learn the properties of the probability distribution using a Restricted Boltzmann Machine (RBM) model [174, 175]. Note that RBM models are techniques taken from deep learning, which is a method utilised in Artificial Intelligence applications (further detailed in Chapter 3). This learning model allows the system to statistically predict projective orchestrations from an input piano score. In *Live Orchestral Piano*, which aims to be a real-time orchestral music generation, the user inputs the piano score via a MIDI keyboard. The program transcribes the MIDI input into a musical score, which is then input into their pre-trained network (using the repertoire of projective orchestration examples) in order to generate a projective orchestration based on the user's piano input. The generated orchestration is directly sonically rendered in the program. Contrarily to the other systems detailed previously, *Live Orchestral Piano* does not try to spectrally match an input file. Rather, it bases its orchestrations on previous knowledge, using machine learning methods to train a probabilistic model that allows the prediction of orchestral combinations from a MIDI input. Furthermore, the timbre mixture emerging from the instrumental combinations is not directly addressed but relies on the composers' work when they

wrote their orchestral pieces. This approach is similar to the content of books on instrumentation, where examples of instruments combinations previously experimented are presented. However, the use of machine learning methods is an interesting approach that differs from the previous attempts for overcoming some of the musical orchestration's complexities.

The field of computer-aided orchestration is still a relatively new area of research. This is perhaps due to the complexity of harnessing every musical orchestration challenges. Nevertheless, the interests for developing computing systems to aid in different aspects of orchestration are growing among the research community, mostly pushed forward by different initiatives that have occurred in the last 15 years, as summarised in Table 2.1. Notwithstanding, the research presented in this thesis does not intend to address musical orchestration nor it is aiming to produce a new computer-aided orchestration system. However, the results of investigations towards identifying techniques for the analysis and control of instrument timbre combinations would have contributions in the field of computer-aided orchestration. For example, harnessing the perceptual qualities of instrument timbre combinations by processing audio files could offer a solution to narrow the large number of solutions that are often generated by such systems.

Systems	Criteria	References
Spectral Music Movement (Various systems, researchers and composers, initiated in the 1970s)	Use of computers and electronic devices to analyse and manipulate spectral information of different types of sound, used as material for multi-instruments composition.	[11, 150, 154, 176]
SPORCH - Psenicka (2003)	Target file matching approach. Instrument database pre-analysed and indexed, use of an iterative matching algorithm by comparing peaks of the target sound and each instrument from the database (return the closest instruments in terms of frequency).	[157]
Rose and Hetrik (2005)	Spectrum analysis of a target sound, proposition of orchestrations approaching the target file using a Singular Value Decomposition (SVD) method.	[156]
Hummel (2005)	Iterative matching algorithm based on spectral envelopes of a target sound and instruments.	[158]
Orchidée - Tardieu and Carpentier (2008)	Target sound analysis, pre-analysed and indexed instruments database. Combinatorial algorithm to match instruments combinations with target sound.	[159, 160]
Ato-ms - Esling (2012)	Generation of orchestrations matching a target sound within a time space, using a multi-objectives and time-series matching algorithm.	[164]
Orchids - Esling and Bouchereau (2014)	Abstract and temporal orchestration of target sounds, using of pre-analysed and indexed audio database. Multi-objectives matching algorithm, informed by acoustical and psychoacoustical information from analysis of the target sound.	[166]
Live Orchestral Piano - Crestel and Esling (2017)	Projective orchestration from MIDI input data. Use of probability distribution models based on analysis of pre-existing projective orchestrations for the suggestions of instrument and notes combinations.	[172]

Table 2.1: Summary of the different approaches used in the development of computer-aided orchestration systems, with their names, criteria, and references for.

2.6 Chapter Discussions

The human's ability to hear sounds involves several mechanisms and processes that are usually effortlessly and quickly performed. First, sounds, which are a succession of vibrations, need to be captured, a process primarily done by the principal component of the human auditory system: the ears. Today's understanding of this organ is the result of a long standing series of investigations, building on works such as research from Fallopio in the sixteenth century [23, 24], Cotugno [25, 26], and Corti, known for describing the organ of Corti [27, 28]. It has been identified that three components, each having different roles, form the human hear: the outer ear, the middle ear, and the inner ear. The outer ear collects and leads the audio signals toward the middle ear, where different mechanisms transform the air movements into liquid movements, which are then transduced by the inner ear into the auditory nerve. This information is then conveyed to the auditory cortex and the auditory association cortex, both located in the temporal lobe of the cerebrum, as illustrated in Figure 2.2, where the sonic information will be processed by the brain. Here, we can see that just the action of capturing the whole information of a sound already involves many mechanisms and processes.

Auditory perception is the ability of processing and interpreting the sonic properties captured by the auditory system. In section 2.3.2, it was identified that four major characteristics of sound can be retrieved: pitch, loudness, duration, and timbre. These sound characteristics provide essential information in detecting the source of the sound and the potential movements of it, contributing to the acquisition of spatial information. Human auditory perception is also capable of discriminating the overall sonic events in order to focus and extract specific sounds. For example, we are able to focus on, and usually understand, the singer in a musical song containing instruments and vocals playing simultaneously. Auditory perception, combined with other cognitive processes such as memory, allows us to understand the information of an audio signal. This process is important for communicating with others and allows us to understand the meaning of a sound. For example, we understand that noises emitted by fire alarms represent danger. Furthermore, we can identify and recognise the source and type of a sound. For examples, we can identify the sound of an instrument or also if it is a voice of a known person. Here, timbre is a key characteristic in the process of identifying and recognising sound sources.

This study does not intend to model musical perception, neither to completely detail and comprehend the different processes of understanding music. However, it focuses on the perception of one specific attribute of sound for musical purposes. A brief introduction of concepts

related to music perception has highlighted the importance of four major characteristics of sound (i.e., pitch, loudness, duration, and timbre). It has been identified that these low-level sonic attributes inform higher-level musical information, which defines how we perceive music. Furthermore, it has been explained that pitch, loudness, and duration can be measured and represented in music, whereas for timbre it is a more challenging task. Nevertheless, this sonic attribute represents several important aspects in experiencing music, especially in regards to perceptual characteristics.

The review of the studies on timbre illustrated that this component of sound has been the subject of many investigations since the early experiments conducted by von Helmholtz in the nineteenth century. These early works suggested that timbre, sometimes referred as tone or sound colour, was all the sonic attributes but pitch, loudness, and duration that allow one to distinguish that two sounds are dissimilar: a rather ambiguous definition that actually illustrates the complexity of this attribute. One layer of complexity is perhaps the different timbral meanings. In regards to the research developments presented in this thesis, there are two paradigms that are most relevant. This study focuses on sounds produced by traditional Western instruments. Thus, the first property is related to instrument timbre, which represents the sonic characteristics that define specific instrument sounds. The second paradigm is related to the use of timbre to represent perceptual properties of the sounds, and within this study, by using verbal descriptors.

The investigations of instrumental timbre have suggested that different properties characterise specific instrumental sounds, resulting from spectral information, such as spectral energy distribution, and temporal information, such as onset and offset characteristics. We can note that timbre is not unidimensional, unlike pitch, loudness, and duration. This was also suggested by the studies investigating the use of verbal descriptors of timbre qualities and identifying their acoustic correlates. Here, several experiments have demonstrated that retrieving specific acoustical properties could inform on perceptual properties of timbre and be represented by words such as *bright*, *dark*, *pure*, and *sharp* for examples. The use of such words from the everyday language to describe sonic qualities instead of using their acoustic correlates is common practice among musicians and composers [177]. Thus, handling verbal descriptors could aid in representing timbral qualities and also in making the tools accessible to a broad audience by alleviating the need to have expertise in acoustics or psychoacoustics.

The review of works on semantic timbre has shown that establishing a list of verbal descriptors and identifying corresponding methods to retrieve the timbre qualities from audio signals

have been the focus of many investigations. In some cases, suggested descriptors may overlap, such as *brightness* and *brilliance* for example, and similarly for their known acoustic correlates. This indicates that it may be difficult to manipulate all verbal descriptors suggested in the literature. Thus, a selection would be required as well as a review of their correspondent acoustic correlates before incorporating this approach into techniques for harnessing instrument timbre and timbral combinations. Furthermore, most of the studies conducted to identify acoustic cues that contribute to the perception of timbre have used sound samples of isolated individual notes as stimuli, which may not accurately represent how we experience music. Indeed, usually music is composed of different instruments playing simultaneously, and sometimes with vocals as well, and therefore this different sonic information may influence the perception of timbral qualities, either for the task of identifying instruments or in perceiving and describing the 'colour' of the sound. Here, the notion of polyphonic timbre appears to be most relevant in sounds emerging from combinations of instruments.

The discussions about investigations on timbre have shown that these studies have mainly focused on monophonic timbre, which was necessary to begin with in order to understand the different concepts represented by timbre. Surprisingly, there have been far less studies on polyphonic timbre, perhaps due to the complex composition of the sound signals. These studies, using different types of stimuli, have produced mixed conclusions, some of which suggest the acoustic cues identified by the experiments on monophonic timbre could be directly applied to polyphonic timbre. Others have suggested that source separation processes are required, and timbre analysis should be performed on a individual source. The lack of agreed methods for polyphonic timbre analysis illustrates that in order to address **RQ1** and harness timbre properties of instrument combinations, an initial study defining and evaluating a timbre analysis method is necessary. This would also provide data for instrument timbre combinations and contribute to the understanding of polyphonic timbre using a different type of stimuli—multiple instruments notes.

Although a significant amount of studies have focused on timbre, whether monophonic or polyphonic, the review of these works has suggested that neither its number of dimensions nor its analysis methods have been universally agreed. This is also the case for its metrics. Unlike pitch and loudness, values retrieved for timbre properties produce various figures and scales that make the interpretation and manipulation of this property difficult. This challenge needs to be addressed in order to be able to answer **RQ2** and to integrate timbre properties in techniques for harnessing sonic qualities emerging from mixtures of instruments. The role

of this attribute is important in instrumental timbre combinations, creating unique sounds and textures. Here, works on polyphonic timbre would benefit the understanding of perceiving multi-instruments timbre, but also the phenomenon of timbre blending, two important aspects of instrument combinations.

The research presented within this thesis does not aim to address musical orchestration. However, some results of the investigations towards developing techniques for the analysis and the control of instrument timbre combinations would contribute in the field of computer-aided orchestration. This areas of research, that looks at using technology and developing computing systems designed to help in various orchestration tasks, has started in the 1970s with the *Spectral Music* movement. However, the main contributions for developing completely automated computer-aided orchestration systems have been initiated in the 2000s, as shown in Table 2.1. One of the most advanced and complete computer-aided orchestration systems is perhaps *Orchids*, developed by IRCAM. This software proposes to output orchestrations that match the spectral and spectrotemporal content of a target file. *Orchids* generates interesting solutions for very diverse types of targets, however, the number of instrument combinations output for one file can be very large. While this multitude of solutions can be interesting in some compositional processes, the instrument combinations can sometimes be musically very different, which requires an additional listening task to find a solution that matches specific sonic qualities. Therefore, a option to narrow the solutions output by computer-aided orchestration systems focusing on harnessing timbral qualities of instrument combinations could be a solution to overcome a tedious and time-consuming listening task.

It has been discussed above that timbre encompasses various acoustical and psychoacoustical properties that are not necessary explicit for a broad audience. Therefore, using words to describe sonic characteristics, which is a practice widely utilised in various musical tasks [169], could alleviate the need for expertise in acoustics and psychoacoustics for representing timbral properties. This would aid in representing and utilising this sonic attribute. Furthermore, an approach using semantics to describe the desired final sound quality as the target of the instruments combinations search could be an interesting method to overcome the large combinatorial space challenge created by the number of instruments that would compose the musical ensemble. Here, performing various signal processing on audio recordings of instrument notes could enable the analysis and understanding of instrument timbre and timbral combinations. It would help in developing techniques for harnessing such characteristics in methods for controlling the perceptual qualities that would be produced by combining specific instruments.

2.7 Chapter Summary

This chapter has presented background information on a number of different concepts related to the research developments presented in this thesis. First, it described the human auditory system, which allows us to capture, transduce, and convey the sonic events from the ears to the brain, where information is further processed. Auditory perception, which is the ability of processing and interpreting the sonic information captured by the auditory system, was subsequently explained, along with a focus on the processes related to the perception of musical information. The purpose of this section was to describe and understand the different sonic components that are captured and processed by humans and also to identify the important musical cues experienced by listeners.

The discussions about the human auditory perception suggested that the attribute of sound timbre has an important role in the perception of music. Thus, the chapter moved on to define this attribute, along with details on different characteristics that timbre can represent, identifying the most relevant to this study. Then, it presented an overview of the large amount of research on timbre, which was primarily focused on studying monophonic timbre, in order to understand the properties of this multidimensional attribute. This section also listed the studies on polyphonic timbre, which, surprisingly, received less focus from the research community. An analysis of the works on timbre highlighted that polyphonic timbre is most relevant to this study, which looks at harnessing timbral characteristics of sounds emerging from combining instruments.

This chapter continued with a review of the field of computer-aided orchestration, which looks at developing systems to aid diverse tasks in orchestral composition. The overview of the different developments, which have mainly occurred in the last 15 years, have informed some developments that will be presented in Chapters 4, 5, and 6. Furthermore, results of this research could contribute in proposing approaches for narrowing the large number of solutions that are often output by such systems.

Discussions about the information presented in this chapter highlighted that the ability to process sonic information, which is usually effortlessly and quickly performed by the human brain, is actually the result of a complex auditory system and various cognitive processes. The analysis of studies on timbre indicated that this component of sound represents several characteristics, and, although a significant amount of studies have focused on this attribute, neither its number of dimensions nor its analysis methods have been universally agreed. Furthermore,

it was explained that the few studies that investigated polyphonic timbre had produced mixed conclusions, which indicates a lack of agreed methods for polyphonic timbre analysis. It was also highlighted that timbre is an essential component in the sonic mixture emerging from the instrumental timbre combinations. Then, the discussions explained the rationale for harnessing verbal descriptors of timbre characteristics in techniques for the analysis and control of instrument timbre and timbral combinations.

In summary, this background information chapter has delimited the scope of this study and defined a number of notions underlying the research developments that will be presented in the next chapters. Furthermore, this chapter has highlighted some key areas that require further investigation:

- Selection of verbal descriptors of timbre qualities.
- Identification of the acoustic properties correlating with the verbal descriptors.
- Methods to analyse polyphonic timbres.
- Methods to refine the instrument combination space.

3

Artificial Intelligence and Musical Applications

3.1 Chapter Overview

Chapter 2 provided background information on audio perception, timbre, and computer-aided orchestration. These concepts are essential in understanding the scope of this study. However, some of the developments presented in this thesis also rely on approaches from the field of Artificial Intelligence (AI). Thus, the purpose of this chapter is to introduce some important concepts of this field and identify their uses for musical applications.

The first section proposes a brief introduction to Artificial Intelligence. The aim of this introduction is not to discuss and review the whole field of AI. Rather, it will provide information related to the approaches that are most relevant to the developments presented in Chapters 5 and 6. Thus, the section starts with discussing the notion of intelligence and identifying the concept of machine intelligence. Then, different computational representations of intelligence are detailed, with an emphasis on the methods that have been utilised in this research.

Following on, Section 3.4 lists some of the uses of AI in music. This section reviews different works in relation to three musical applications that are most relevant to this study: sound synthesis, music production, and composition. Other musical applications such as music performance and consumption, in which AI approaches have also been used, are beyond the scope of this research, and therefore, are not discussed here. This section aims to highlight the relevance of harnessing AI methods for musical applications. The chapter then concludes with discussions in regards to the concepts and ends with a summary.

3. ARTIFICIAL INTELLIGENCE AND MUSICAL APPLICATIONS

The structure of this chapter is as follows:

- 3.2 - Introduction
- 3.3 - Artificial Intelligence
- 3.4 - Artificial Intelligence Applications for Music
- 3.5 - Chapter Discussions
- 3.6 - Chapter Summary

3.2 Introduction

Artificial Intelligence (AI) is a concept that has inspired many science-fiction authors [178]. However, most importantly, AI is a very active field of research that have produced revolutionary methods for many applications, which started with a workshop at Dartmouth College in 1956 [178, 179]. Research in this area began with the idea that "*every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*" [180, p. 12]. This concept of general AI, or global AI, which consists of computing systems that would have all of human intelligence characteristics, senses, and can reason, is still not reached yet. Nevertheless, since the early experiments, the field of AI research has produced numerous 'intelligent' methods designed to achieve specific tasks. With advances in technology, such as the increase of computer power, storage, and data, combined with a growing research interest and funding, AI has had an important impact in almost all domains of our life. AI now has applications in healthcare [13], with medical diagnosis [181] for example, in industry, such as autonomous cars [182], or also for security, with facial recognition [12] for example. A review of all the applications of AI is beyond the scope of this study, and it would produce an endless list. However, a discussion of its goals and approaches, and most importantly a review of AI methods that could help in harnessing timbre perception of instrument combinations would help in understanding some of the choices underlying the developments presented in Chapters 5 and 6.

It is undeniable that advances in technology, in particular computing technology, have had a big impact in music. Unsurprisingly, AI has also reached different aspects of the musical domain. Several applications of AI into music are discussed in the second part of this chapter, which is used to illustrate the general AI approaches introduced in Section 3.3. The review of works that have used AI for sound synthesis, production, or composition, three musical practices in which parts of this study contribute, will highlight methods that can also be implemented for instrument timbre and timbral combinations.

3.3 Artificial Intelligence

This section proposes to review the concept of Artificial Intelligence, its goals, approaches, and methods. First, it is necessary to consider what (human) intelligence is before defining one that is artificial. This discussion does not intend to be philosophical, nor to argue the concept of intelligence, but rather to give an understanding of the motivations behind harnessing this concept into computing systems. Following the establishment of the human intelligence and its theories, the discussions move on to define the notion of machine intelligence. Then, the last part of this section proposes a review of the different goals that motivate creating AI algorithms. The different methods that have been developed to achieve these goals are listed, along with some applications, and references are provided for further details.

3.3.1 Definitions

Nowadays, the term Artificial Intelligence (AI) is everywhere and sometimes used to group many different applications. As stated previously, the concept of AI takes inspiration from human intelligence. It aims to imitate and simulate different cognitive functions. Therefore, before discussing key concepts of the field of AI research, it may be suitable to distinguish the features of human intelligence, and to review different theories of this concept, which are the foundations of AI. The discussion of these notions will help to define the concept of machine intelligence, or in other words, defining the idea of artificial intelligence.

3.3.1.1 What is Human Intelligence?

Providing an answer to the question 'what is intelligence?' is a real challenge. This notion has been debated since the ancient Greek period where philosophers such as Aristotle formulated a set of defined principles that underlay the human rationality [179, p. 6]. Here, it is suggested that rationality and reasoning may form human intelligence. Since these early formulations, many works have investigated this concept. A historical review can be found in [179, Section 1.2, p. 5–16]. These works led to today's understanding of this notion. However, it has produced many definitions. For instance, Legg and Hutter [183] list over 70 different definitions of the notion of intelligence, which illustrates the difficulty of proposing a single definition, but also the complexity of this concept. It also shows the difficulties of harnessing this aspect into 'machines', or in other words, into computing systems. Nevertheless, one definition that may summarise this notion could be: "*intelligence is a very general mental capability that, among*

other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" [184]. With this definition, we can note that intelligence might involve many mechanisms. Furthermore, the complexity to produce a single definition of intelligence is perhaps emphasised by the fact that there is not only one intelligence, as it will be discussed in the following section.

3.3.1.2 Theories of Intelligence

Information about theories of intelligence can be found in [185] and [186], both focussing mainly on contemporary theories. The discussion of some of these contemporary theories also serves to highlight the different aspects that need to be harnessed in order to model human intelligence.

Thurstone suggests that intelligence is the result of different mental abilities [187]. Many of the contemporary theories of intelligence are based on this approach, notably the ones proposed by Gardner [188], further detailed in the next paragraph, and Carroll [189]. Thurstone based his approach on an analysis of 56 different tests of mental abilities, and proposed that seven primary mental abilities form the human intelligence:

- *Verbal comprehension*: the ability to understand verbal information.
- *Verbal fluency*: the ability to create and use words to produce sentences in order to convey information.
- *Number*: the ability to use numbers and perform arithmetic.
- *Perceptual speed*: the ability to rapidly recognise letters and numbers.
- *Inductive reasoning*: the ability to make a general conclusion from specific information.
- *Associative memory*: the ability to remember the link between unrelated items.
- *Spatial visualization*: the ability to visualise shapes, rotations of objects, identify shapes that can fit together in puzzles.

Here, we can note that human intelligence involves diverse mental abilities. Therefore, to create an artificial intelligence that follows this theory, a computing system would require harnessing and combining all of them.

Gardner, who was influenced by Thurstone's theory, proposed that intelligence was not a single component, but rather the existence of multiple intelligences. In his *Theory of Multiple Intelligences* (MI Theory), Gardner identified eight different intelligences [188, 190], which are as follows:

- *Linguistic*: the ability to read and write.
- *Logical-mathematical*: the ability to solve mathematical problems and to derive a logical proof.
- *Spatial*: the ability to visualise, maneuver, and operate the spatial environment.
- *Musical*: the ability to sing or compose musical pieces.
- *Bodily-kinesthetic*: the ability to move and use the body, or part of it, to perform specific task, such as dancing.
- *Interpersonal*: the ability to understand and interact with other people.
- *Intrapersonal*: the ability to the understand oneself.
- *Naturalist*: the ability to recognise and understand patterns in nature.

We can see that Gardner's MI Theory proposes intelligences that are similar to some of the primary mental abilities proposed by Thurstone. Based on Gardner's theory, some abilities might be more challenging to harness into computing systems (e.g., intrapersonal intelligence) than others. It is also interesting to see that Gardner recognises a musical intelligence.

Another theory is the *Successful Intelligence*, or also named the triarchic theory of intelligence, which was proposed by Sternberg [191, 192, 193]. He defines this notion as the "*mental activity directed toward purposive adaptation to, selection and shaping of, real-world environments relevant to one's life*" [194, p. 45]. Sternberg suggests that the combination of three important abilities contributes to a successful intelligence. The three components are as follows:

- *Analytical intelligence*: the ability to analyse information, evaluate, and solve problems.
- *Creative intelligence*: the ability to generate ideas and options to solve a problem.
- *Practical intelligence*: the ability to carry out options to address a problem, using past experiences and current skills.

According to Sternberg, analytical, creative, and practical abilities form the human intelligence. We can also note that intelligence is regarded as handling problems.

There are other significant theories of intelligence, however, the objective of this part is to illustrate the complexity of this notion, not to propose a comprehensive review of all the theories. Here, we can observe that intelligence is not a single element, but rather involves diverse components and abilities. These theories provide a framework towards modelling this concept. They also provide measurement methods, which helps in understanding and identifying aspects of intelligence. However, to computationally simulate human intelligence, harnessing all the different mechanisms would be required.

3.3.1.3 Intelligent Machines

Following the idea of artificial intelligence put forward during the 1956 Dartmouth AI workshop (i.e. simulate every aspects of the notion), and the discussions of the different theories of intelligence have shown that in order to create a machine that encompasses human intelligence, several different mechanisms need to be harnessed. However, building on the idea of multiple intelligences, shared by many contemporary theories, a machine could be considered as intelligent if it succeeds to simulate one or more intellectual competences. Since the early experiments in the 1950s and 1960s, the field of AI research has produced numerous methods that have been successful in imitating aspects of human intelligence to achieve specific tasks, aided by advances in technology. Some of these methods will be discussed in the next part of this section.

3.3.2 Computational Representations of Intelligence

Similar to the notion of human intelligence, which regroups diverse abilities as discussed in the previous part, the term Artificial Intelligence regroups different concepts. Since its creation in the 1950s, this field of research has created numerous methods to harness different aspects of intelligence and has produced solutions to address many problems. Nowadays, AI has applications in almost any aspect of our life. This section proposes a brief review of some AI goals, applications, and methods. A comprehensive description of the different AI approaches, goals, and methods can be found in [179].

3.3.2.1 Approaches and Goals

Research in AI tries to achieve different goals, which are based on the different approaches representing diverse cognitive abilities exhibited in human intelligence. Some of these AI goals are discussed below:

Reasoning and Problem Solving: The search of developing algorithms that can reason and solve problems is perhaps one of the first goals explored by AI works. The model is based on the step-by-step reasoning method used by humans and thus, intelligent systems incorporate algorithms that follow the 'formulate, search, execute' model. Here, the goal of the intelligent system is to output a solution that solves a problem. Thus, the research aiming to achieve this goal focuses generally on developing methods to optimise search algorithms.

Knowledge Representation: Many goals that AI aims to achieve require data that represent knowledge. This is particularly true in problem-solving for example. Thus, knowledge representation is an essential area of research in AI. The focus here is to find ways to translate the knowledge, which can sometimes be very large and diverse, into data that can be manipulated by AI algorithms. The creation of representations of knowledge is called ontological engineering.

Planning: Planning is a core part of AI concepts. The intelligent systems need to be able to devise a plan of operations to achieve the defined goals. Therefore, systems have to predict the consequence of their actions and identify those that would maximise the success of achieving their goals. If the system has to interact with factors other than its own actions, such as air transport where there are multiple planes flying, for example, the intelligent system needs to assess and adapts its predictions and actions in regards to all factors.

Natural Language Processing: An attribute that is uniquely part of human intelligence is the ability for language, which is an important component of the Turing Test [195]. Communication is based on speaking and writing. Naturally, research in AI also aims to harness human languages. The two main reasons to incorporate this aspect into computing systems are, first, the ability to interact with humans and, secondly, to be able to acquire knowledge, especially from written language. This is of interest for information retrieval and automatic translation

tasks for examples.

Motion and Manipulation: The ability to make specific movements and the manipulation of objects have been suggested by some theories as part of the human intelligence. Harnessing these two aspects are also part of the AI goals. Advances in this domain are important for the field of robotics, for instance. Here, intelligent systems are required to allow robots to (successfully) perform different tasks, such as navigation and object manipulation. Computing systems need to be able to assess and learn from their surrounding environments to inform and execute the different tasks.

Perception: As some theories suggest, perception is also part of human intelligence. It is defined as the ability to organise, identify, and interpret the information provided by our sensory system in order to represent and understand our environment. For example, auditory perception, which is related to this study, has been discussed in Chapter 2 of this thesis. Further details about human perceptual abilities can be found in [21, *Part I - Perception*]. Harnessing perception into computing systems is also one of the AI goals. Here, the sensory system is represented with sensors (such as cameras and microphones). Intelligent systems would, therefore, need to be able to use the information provided by such sensors. This aspect is important in facial recognition and speech recognition processes, for example.

Creativity: As discussed in Section 3.3.1.2, there is a creative intelligence, notably put forward by Sternberg in his *Successful Intelligence* theory. Some AI research groups aim to harness creativity into computing systems. Some research tries to develop systems capable of recreating human creativity, other tries to enhance human creativity by developing systems to achieve creative tasks. Research in artificial creativity also aims at better understanding human creativity [196]. Advances in this area find applications in artistic tasks. For instance, there is a research community investigating artificial musical creativity [197].

Learning: The ability to learn is an essential part of intelligence. This process serves to acquire knowledge, which leads to an improvement of performance. Unsurprisingly, learning is also a core aspect of AI. Research in this area forms the very active field of *machine learning*. Here, the aim is to improve algorithms automatically without the task of having to (re)program

them each time they encounter new information. Thus, research in machine learning is looking at developing models that can learn from data, but also make predictions based on data previously seen. Learning algorithms are sometimes used to address tasks that are difficulty programmable, such as image recognition. Intelligent systems are learning if their performances improve with assessing new data. Machine learning finds applications in many areas, such as facial recognition, prediction of financial markets, and email filtering (such as detection of spam emails). Some machine learning methods have been used in research developments presented in Chapters 5 and 6.

The overall goal shared by some AI research groups is to harness all the different abilities exhibited by human intelligence into a single computing system, which is referred to as artificial general intelligence. Some believe that such a task will be eventually possible, arguing that machines would be able to exceed human abilities at some point [198, 199]. However, such machines would require combining and executing all of human abilities simultaneously, which is still a challenge and debated among the research community and philosophers. Nevertheless, research in each AI goal has produced numerous solutions and methods to solve specific problems. Some of the methods produced by the field of AI research will be discussed in the following part.

3.3.2.2 Methods

The previous part delivered a discussion of different goals that the field of AI research aims to achieve. Since its creation in the early 1950s, this field of research has produced numerous methods to solve specific problems in an 'intelligent' manner. A comprehensive review of all AI methods is beyond the scope of this study. Thus, only a selection of methods related to some research developments presented in this thesis will be discussed in this section. Please note that specific AI methods that have been used in some developments of this study will be further explained in Chapters 5 and 6.

Search algorithms are used for many tasks, especially for problem solving. The simplest way to solve a problem may be to test all the possible options. In the case of route planing, for example, the algorithm will test all the possible routes and return the best option to solve the problem. While this approach works for small search space problems, it can quickly become computationally expensive to try all options. Therefore, some search algorithms use heuristics to reduce the search space [200], others use mathematical optimisation to refine the search

space. Another search method is evolutionary algorithms [201], which is inspired by biological evolution. See [179, *Part II - Problem-solving*] for further details about different search methods.

Research in AI has also produced several methods to classify data, which has an important role in machine learning. Here, the objective of these methods is to automatically decide in which category an object would fit, based on the learning ability. Different classifiers have been developed, such as Support Vector Machines (SVM) [202] and Artificial Neural Networks (ANNs) [203, 204], to name but two. The performance of classifiers vary upon the type of data, the problem to solve, and the training methods. Nevertheless, these methods have been used in many machine learning approaches to achieve different tasks, such as object recognition, Internet search engines, and speech recognition, for examples. Different classifiers have been used in the developments presented in Chapter 5.

Several other methods have been put forward by the field of AI research and are applied in almost every aspects of our life. The next section will continue to discuss some AI methods, with a focus on their applications in the musical domain, to illustrate the potential of AI techniques for musical purposes.

3.4 Artificial Intelligence Applications in Music

The previous section has described the notion of intelligence and presented an overview of the field of Artificial Intelligence with discussions about its goals, approaches, and methods. This section aims to illustrate the implications of AI in the musical domain. Thus, some uses of AI in different musical contexts are reviewed. The discussions focus mainly on three musical practices: sound synthesis, music production, and composition. Other practices, such as music performance, are not directly related to this study and its outcomes, therefore, a review of works related to such practices will not provide information that supports the understanding of the research presented in this thesis.

3.4.1 Artificial Intelligence and Sound Synthesis

The first review of the use of AI for musical practices focuses on applications in sound synthesis. While this study does not aim to address the challenges of sound synthesis, neither does it have direct outcomes for such practice, some of its developments rely on the use of audio samples and their combination. Therefore, it may be interesting to review a selection of works that have used AI for sound synthesis.

Concatenative sound synthesis (CSS) is perhaps one of the first synthesis techniques that has used AI methods [205]. CSS produces new sounds from combining smaller audio files. Here, AI techniques have been applied in the matching algorithms that try to find and combine audio samples within a sound collection. For example, rule-based models and data-driven models, where the matching algorithms respectively follow rules based on human expert knowledge or are based on the data itself, have been used in CSS systems [205, 206].

AI techniques have also been applied in programming sound synthesisers. For example, Miranda used two machine learning techniques, namely inductive learning and supervised deductive learning, in his system ARTIST (an acronym for Artificial Intelligence-aided Synthesis Tool) [207]. These techniques have been utilised to learn programming synthesisers' parameters according to high-level properties, such as qualitative descriptions. For instance, the user can ask to produce a sound similar to a bell and the system will find the parameters to produce a bell-like sound. Another example of programming synthesiser using AI is *midimutant*[1], an application that runs on a Raspberry PI[2] and aims to program synthesisers using artificial

[1] `https://fo.am/midimutant`
[2] `www.raspberrypi.org`

evolution, which means it starts by generating a random patch and evolves it to obtain sounds that match the user's ideas. This program has been recently developed by FoAM, a network of transdisciplinary research labs, in collaboration with Aphex Twin, a British electronic musician, and is based on work by Horner [208], which used genetic algorithms for Frequency Modulation (FM) synthesis.

Researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL)[1] have developed an algorithm that can synthesize sounds from a silent video [209]. Here, they have trained a neural network algorithm with videos they recorded (containing 'real' sounds), in order to learn the sonic content of the videos, and thus, be able to predict the sound features of videos. From the prediction of sound features, the algorithm then generates a waveform, using an example-based synthesis technique. Here, the researchers have used Artificial Neural Networks (ANNs), a machine learning technique, to predict sound features, which inform the concatenative sound synthesis technique.

ANNs have been extensively used in the field of machine learning for many applications. While such techniques have been proven to be successful, the training of a system implementing ANNs can be computationally intensive. Thus, some researchers have developed ANNs that can train themselves, analysing large amounts of data to learn. This has resulted in the field of deep learning, a sub-field of machine learning [210, 211]. Deep learning methods have been applied in many fields and also in sound synthesis. In 2016, Google DeepMind research group[2] introduced *WaveNet*, a deep neural network capable of generating raw audio waveforms [212]. Here, they have been able to synthesis human-like speech sounds, using different audio recording databases to train their deep learning model. They also applied their model to musical data, using a database of piano recordings, and another one containing mixed music audio recordings.

Another Google research group, Google Brain[3], has recently implemented AI methods for sound synthesis applications. Magenta[4], which is a project from the Google Brain team, has developed *NSynth* (Neural Synthesizer)[5], a sound synthesiser that is based on *WaveNet*'s deep neural networks approach [213]. Here, they created a large dataset containing audio samples of around 300000 notes from nearly 1000 instruments, which serves as training data for their

[1] `http://csail.mit.edu/`
[2] `https://deepmind.com`
[3] `https://research.google.com/teams/brain/`
[4] `https://magenta.tensorflow.org/`
[5] `https://magenta.tensorflow.org/nsynth`

deep learning algorithm. *NSynth* aims to use deep learning and data-driven methods to aid users exploring a sound space that may not be achievable with synthesisers using oscillators and wavetables. For example, its training enables the audio morphing of different instruments sounds—note that the idea of combining sounds from two instruments (or parts of instruments) in sound synthesis method is not new, see [214] for example.

AI methods have been applied to different aspects of sound synthesis. The discussions above have shown that rule-based models, evolutionary computing models, and machine learning methods, among other approaches, have been used whether for the selection of audio samples, for programming sound synthesizers, or for generating audio waveforms. AI advances also find many applications in speech synthesis, briefly mentioned in discussions about Google DeepMind's *WaveNet*. As detailed in Section 3.3.1.2, the ability to communicate, especially with speaking, is part of human intelligence, hence it is expected to see the use of AI methods for speech production, whether for coherency of data to output or for producing human-like sounds. Advances in this domain have produced several general public applications, such as Apple's *Siri* and Amazon's *Alexa* to name but two. Without doubt, intelligent methods developed by the AI research community will find an increasing interest from the sound synthesis field for addressing some of its challenges.

3.4.2 Artificial Intelligence and Music Production

The study presented in this thesis investigates aspects of timbre perception of combinations of instrumental notes. Some developments detailed in Chapters 4 and 5 may have applications in the domain of musical production. The use of intelligent tools for music production tasks has recently increased, and several research groups are investigating the potential use of AI methods in aiding some music production tasks. This section reviews some research aiming to harness an AI approach for music production purposes.

Since the early suggestions on automating music production tasks in the 1970s, such as Dugan's automatic microphone mixing approach [215], many methods, systems, and applications aiming to assist or completely perform such tasks have been developed. For example, Cartwright and Pardo have proposed a system capable of automatically performing equalisations on audio signals based on descriptive attributes input by the user [216]. To achieve this task, they developed *SocialEQ*[1]: a web-based application utilised to learn descriptive attributes

[1] `http://socialeq.org/`

and their correspondent equalisation curves. Here, users can define their notion of the descriptive term they entered in *SocialEQ* by listening and rating different equalisations of an audio file. This action is actually training a machine learning algorithm developed to match equalisation curves to the user's appreciation. Then, *SocialEQ* can automatically perform an equalisation that follows the user's desired sound quality onto the audio file, which can be seen as an automatic selection of equalisation presets. This method also provides a framework to identify which equalisation curves can match a specific sound quality.

In [217] and [218], De Man and Reiss propose a system that can automatically perform some audio mixing tasks. They have used a rule-based model, informed by the literature about audio engineering practices, to perform balance, pan, compression, and equalisation tasks on audio files. Here, the system analyses individual tracks and groups of tracks (e.g., drum, guitars), and incrementally decides which task to perform on each input, following a rule list. De Man and Reiss performed perceptual experiments to evaluate the performances of their autonomous mixing system. Their results show that the mixes produced by their system have no significant difference from human mixes, which suggests that expert systems, based on knowledge and rules, might be capable of performing some audio mixing tasks at human level. However, such systems have not (yet?) replaced professional sound engineers in mixing complete musical pieces.

The last decade has seen many works focusing on developing intelligent methods and systems for music production purposes. Such systems have been able to perform various mixing tasks, such as equalisation, level-balancing, panning, and compression, and also applying effects on audio signals [219]. The majority of these systems have used knowledge and rule-based models, following audio engineering practices and examples. However, knowledge and rule-based models may not encompass enough information to perform every music production task at human level. In [220], Wilson and Fazenda suggest adding an evolutionary computing approach to expert systems, mentioning the lack of human perception criteria in such system, which could explain that intelligent systems do not out-perform human-made mixes. Music production tasks rely on scientific knowledge and methods. However, the result mainly depends on the audio engineers and listeners ears. Perhaps, advances in AI research focusing on harnessing human perception abilities could offer a way to incorporate audio engineers' subjectivity and creativity into intelligent music production tools.

3.4.3 Artificial Intelligence and Composition

The musical practice that may have had the most applications of AI methods is perhaps composition. Since the early developments in the 1950s, several AI techniques have been used, especially in algorithmic composition [221]. There are two different approaches on how to use AI for musical composition. The first point of view is to use AI techniques to aid composition, which is part of the computer-aided musical composition field. The second approach is applying AI for automatic composition, where intelligent systems generate the whole musical piece without human input. This section will review a selection of works that use AI methods, with a focus on the most recent developments, in order to highlight the benefits of AI for this musical practice. A comprehensive survey on AI and algorithmic composition can be found in [221].

3.4.3.1 Intelligent Companionships

This part reviews some works that aim to harness AI methods for aiding musical compositions. Here, the goal of these methods is not to generate a complete musical piece, but rather to propose raw materials for composers, or to assist in some part of the compositional process.

One of the first experiments investigating the use of AI methods for music has been conducted by Hiller and Isaacson, in the 1950s, which resulted in the composition of the famous piece entitled *Illiac Suite* [222]. Hiller and Isaacson have experimented with rule-based systems and Markov chains (a stochastic process [223]) applied to formal music composition. In this case, the *Illiac Suite* piece has been completely generated by a Markov chains algorithm. However, this basic machine learning approach, which has inspired many researchers and composers since the early experiments, has been considered as a source for compositional materials rather than an approach to generate complete compositions, due to its limitations for music—low-n orders Markov chains produce random data, while high-n orders tend towards producing data very close to the training corpus [224]. For example, Tipei applied Markov chain process on melody information in his program *MP1* [225], while Pachet utilised variable orders Markov chains for his interactive improviser *The Continuator* [226]. Davismoon and Eccles applied Markov processes for melody and rhythm [227].

Artificial Neural Networks (ANNs) [204] are algorithms inspired by a human neuron network, which consists of interconnected sets of artificial neurons that can receive, process, and send information within the network. ANNs have typically been implemented as a machine

learning method, using set of examples to train the network to identify and recognise patterns. One of the first implementations of ANNs for musical composition was conducted by Todd, where he used three-layered recurrent ANNs to produce melodies based on monophonic melody examples [228]. Since, ANNs have also been used for different musical purposes such as harmonisation [229], jazz improvisation [230], and poly-rhythmic structures [231], for example. ANN methods have been, for the most part, utilised for learning melodic and rhythmic patterns from input examples in order to suggest new musical materials, which is a common practice in algorithmic composition.

The recent developments of deep learning methods have also been applied to compositional tasks. For example, a collaboration between Yotam Mann and the Google Magenta and Creative Lab teams have created *A.I. Duet*[1], a system that interactively responds to a piano input that can be played from a MIDI keyboard. The system contains six different task-specific previously trained recurrent neural networks that allow the generation of responses based on the user input. Here, the system aims to interact with the users and to propose outputs that can expend their creativity by suggesting new musical material.

In 2011, IBM announced their intelligent system named Watson [232], which was originally developed to compete in the American quiz show *Jeopardy!*. Recently, some IBM research groups have been investigating the use of Watson's technology for musical purposes, which resulted in *Watson Beats*[2]. This system is designed to generate music from minimum input material, as few as 15 seconds worth of data. In order to achieve this goal, *Watson Beats* has been trained with knowledge specific to pitch, rhythmic structure, instrumentation, and also musical genre, learnt from millions of musical examples. This training was performed using Restricted Boltzmann Machine (RBM) algorithms, which are stochastic neural networks. The goal of *Watson Beats* is to suggest expanded versions of users' inputs, using deep learning approaches to understand the musical patterns from a large set of examples. Research groups behind *Watson Beats* aim to understand the musical composition's mechanisms and expend composers' creativity with such intelligent systems.

A similar approach of using deep learning methods can be found in *Orb Composer*[3], a software developed by the French company Hexachords which is not yet publicly available. *Orb Composer* aims to help composers' creativity by suggesting instrument combinations,

[1] `https://aiexperiments.withgoogle.com/ai-duet`

[2] `https://www.ibm.com/watson/music`

[3] `http://www.hexachords.com/orb-composer`

melodies, and musical structures based on high-level (such as desired emotions) and low-level (such as instruments or tempo) parameters. Hexachords state on their website that their intelligent system is "*built up from a deep knowledge and analysis of music and a deep understanding of composition process*"[1] based on properties of 1000 different instruments and playing styles, but without specifying further technical details about their deep learning approach. Nevertheless, the goal of *Orb Composer* is to interact with composers by enabling them to modify and adapt the material suggested by this intelligent system.

Since the early experiments, advances in the field of AI have been applied for compositional purposes, especially in the field of algorithmic composition. Rule-based and machine learning methods have been the dominant approaches implemented in compositional systems, usually aiming to identify patterns, rules, and composers' style from previous existing musical pieces and composition techniques. The recent developments follow the trend of implementing deep learning methods to harness the large knowledge and available musical data in order to propose interactive tools for composers, with a back-and-forth information flow. Deep learning approaches for artificial musical creativity purposes may be the route to take to develop intelligent musical companionships.

3.4.3.2 Intelligent Artificial Composer

Research on harnessing AI methods for composition has also focused on developing systems capable of automatically composing entire pieces of music. The goal here is to create an intelligent artificial composer, which can compose music with almost no human input. This section reviews systems that try to encompass, imitate, and extend composers' skills to generate music autonomously.

The idea of creating a computing system capable of generating entire piece of music has been the goal of many research groups. For instance, the experiments conducted by Hiller and Isaacson, mentioned in the previous section, resulted in the generation of the piece *Illiac Suite* in 1956 [222]. Here, they based their approach on a combination of rule-based models and Markov chains, which were designed to analyse input data to statistically identify patterns, which was then used to generate new data. This approach of using statistical models to identify patterns from examples has been utilised in several works. However, as mentioned in the previous section, they were mainly used as materials to work on for composers.

[1] `http://www.hexachords.com/orb-composer/`

In the 1980s, Cope started working on developing computing systems capable of analysing and learning mechanisms of composers' style. His research resulted in the creation of the software *EMI* (short for Experiments in Musical Intelligence), which can generate musical pieces in the style of a specific composer based on an analysis of previous pieces [233]. Cope used Augmented Transition Network (ATN)—a graph theoretic method that is usually utilised in natural language processing to analyse the structure of sentences—for the analysis of musical structures and short sequences. *EMI* used a combination of music analysis methods and recomposition of short musical sequences to generate complete pieces, which produced impressive results. Cope continued his research on musical intelligence and developed the *Emily Howell* program, derived from *EMI* [234]. While *EMI* was trying to copy a composer's style, the *Emily Howell* system aims to generate musical pieces in its own original style. Its approach is based on a trial and error rating model, where the user provides feedback on the material output by the system, which is used to guide the generation of new musical materials. Cope's *Emily Howell* system has generated musical pieces for two commercialised albums: *From Darkness, Light* (2009) and *Breathless* (2012).

A research group from the University of Malaga and *Melomics*[1] [235] have developed two systems capable of generating entire pieces of music without almost no human input. Their first system, named *Iamus*[2], is a computer cluster that runs evolutionary algorithms to autonomously generate full pieces of contemporary classical music using only a list of instruments and a preferred duration defined by the user [236]. Some musical pieces composed by this system have been performed by the London Symphony Orchestra and recorded for the production of an album entitled *Iamus*, released in 2012. Their second system, *Melomics109*[3], based on the same approach as *Iamus* but with larger computing capabilities, was developed to generate musical pieces in a variety of different musical styles. *Melomics109* has created a repository of over 1 billion generated songs. While these two systems have produced impressive results, the technical details of their generative models are not explicit, which makes their approach difficult to evaluate.

Since its creation in 1997, the Sony Computer Science Laboratories (CSL) music team[4], led by Pachet, who developed *The Continuator* mentioned in the previous section, has been investigating interactivity and creativity in music, and their integration into computing systems.

[1] `http://melomics.com`
[2] `http://melomics.com/iamus`
[3] `http://melomics.com/melomics109`
[4] `https://www.csl.sony.fr/music.php`

Their recent *Flow Machines* project[1] has been focusing on developing an intelligent system capable of generating music autonomously [237, 238]. Their system integrates a combination of machine learning methods for extraction of musical features and learning styles, combinatorial optimisation methods to comply to styles' constraints, and knowledge representation methods to represent musical properties. Here, the user can just select the style and duration of the song, and *Flow Machines* generates an original composition. For instance, *Flow Machines* has been used to compose *Daddy's Car*[2], a pop song composed in the style of The Beatles, with some human input for the final product (arrangements, production, and lyrics by French composer Benoît Carré). *Flow Machines* has also been utilised to recompose Beethoven's *Ode to Joy* in seven other styles, including bossa nova and jazz, to name but two [239]. It can also automatically generate chorales that follow Bach's characteristics, using deep learning methods [240]. The twenty years of research behind *Flow Machines* has produced a system capable of automatically generating impressive musical pieces.

A Luxembourgian start-up, Aiva Technologies[3], has been developing an intelligent system for automatic classical and symphonic music composition. In 2016, they created a system called *AIVA* (short for *Artificial Intelligence Virtual Artist*), designed to compose music for movies, commercials, games, and trailers. There are not many technical details about their system except that it uses a deep learning model using NVIDIA's cuDNN[4] and TensorFlow libraries[5], trained using analysis of music partitions from composers such as Mozart, Beethoven, and Bach. Nevertheless, *AIVA* is the first non-human composer recognised by a music society—in this case by the SACEM (Société des Auteurs, Compositeurs et Éditeurs de Musique), a French author's rights organisation. For instance, *AIVA* has composed a 24 tracks album entitled *Genesis*, released in November 2016. Similar to previous automatic composition systems, *AIVA* is trained with a large set of musical pieces written by human composers, a process that aims to identify the mechanisms of musical composition in order to generate new compositions.

The quest for developing an intelligent artificial composer has been the subject of many research groups for several decades. With advances in AI research, technology capacities, and data available, researchers have been training systems to learn composition mechanisms and composers' styles, from large sets of examples. The first intelligent systems were generating

[1] `www.flow-machines.com`

[2] `https://soundcloud.com/user-547260463/daddys-car`

[3] `www.aiva.ai`

[4] `https://developer.nvidia.com/cudnn`

[5] `www.tensorflow.org`

new compositions based on a composer's style, nowadays, the trend is to generate original pieces that aim to match human standards. The use of AI for automatic musical composition is perhaps the application that is the most controversial among the music community, but at the same time, is the approach that intrigues and interests the general audience and draws attention in the media [1].

[1] For example, `www.flow-machines.com/tag/flow-machines-press` groups the significant media attention about the Sony CSL's *Flow Machines* project.

3.5 Chapter Discussions

This chapter has highlighted that human intelligence is a complex notion that actually involves different cognitive abilities. The idea of harnessing every aspect of intelligence into computing systems, the fundamental of the field of Artificial Intelligence created in the 1950s, still remains a challenge. However, this very active field of research has produced numerous methods to address specific problems in almost any fields. Nowadays, AI has applications in many aspects of our life, such as for healthcare, for autonomous cars, and also for security.

Unsurprisingly, AI has also been applied in the musical domain. The discussions about its use in audio synthesis have shown that AI methods can be applied to different audio synthesis questions, such as aiding to program synthesizers' parameters, predicting and creating sounds for videos, and also in developing interactive synthesizers. These applications of AI methods present interesting pathways for addressing some of the audio synthesis challenges, however, in regards to this study the most interesting approach is the development of search-based methods that have been successfully applied to audio sample selection in concatenative synthesis frameworks. Such methods will require further investigations in order to define their potential applications in algorithms designed to search for combinations of instrumental notes using audio samples.

This chapter also discussed the application of AI methods in some music production tasks. Most of the research consist of automating different audio mixing tasks, such as equalisation and level-balancing. They usually used a combination of rule-based models and machine learning methods to apply the scientific knowledge about music production, but also to create templates that could be automatically applied to any signals. It has been shown that such systems have been able to perform successfully specific tasks. However, these methods do not completely harness audio engineers' abilities as they can lack the perceptual analysis and evaluation of the audio engineers when performing such tasks.

The musical practice that has seen the most applications of AI methods since the 1950s is composition. The review of works investigating the use of AI methods for musical composition has suggested two different approaches. One of these two aims to create systems capable of automatically generating complete pieces of music that can match human quality. Most of the research in this area has used combinations of rule-based models, evolutionary computing models, and machine learning methods to train systems using pieces written by human composers and compositional rules also defined by a human. Here, AI methods are used to identify

musical and compositional characteristics from a set of examples, aiming to copy and simulate composers' abilities. The recent advances in this area, such as *Flow Machines* and *AIVA* have taken advantage of the developments of deep learning methods, with the use of large datasets, and have produced impressive results. Such systems can generate thousands of songs with only a few defined parameters. However, the generations mainly depend on the training examples and may result in generating similar materials. A comprehensive survey about the use of deep learning methods for generating musical content can be found in [241].

The second approach consists of developing methods and systems that can assist composers in different tasks. Here, AI methods are used for the development of interactive tools, offering a companionship to the users, with an aim to aid in their musical tasks. AI approaches have been applied in algorithmic composition using AI methods to generate material for composers to work on. The recent developments aim to be interactive, offering an interface between composers and machines, with the hope of assisting composers' creativity. This study aims to aid composers in their metier by investigating and identifying methods for the analysis and control of instrument timbre and timbral combinations. Thus, AI methods systems designed to assist composers may provide useful information for processing and manipulating data related to perceptual qualities produced by combining instrument notes in computing systems.

As it has been discussed in Section 3.3.2.1, the field of AI aims to achieve different goals. Methods developed to achieve some of these goals would apply to this study. For instance, problem solving with a focus on search algorithms could aid in finding the right instrument combination from the infinite potential combinations. Methods from knowledge representation would help to find a way to represent the perception of timbre from sonic sources into computer's data. Finally, the discussions about the use of AI in music have highlighted that machine learning methods have been applied in many systems, whether to learn musical rules or composers' styles. Such methods would aid in linking search algorithms and knowledge-based models in order to harness timbre properties of instrument note combinations.

3.6 Chapter Summary

This chapter has presented background information on the concept of AI, which has informed some of the research developments presented in this thesis. First, it proposed a brief introduction of the field of AI. It started with discussing the notion of human intelligence. Here, the discussions about different theories of intelligence have highlighted that human intelligence is a complex notion that involves several cognitive abilities, therefore, there is more than one intelligence. Then, the discussions continued with defining the notion of AI and reviewing some of its approaches and goals. Similar to the different human intelligences or abilities, the field of AI research aims to achieve different goals, which can sometimes overlap. Nevertheless, research in AI has produced several methods to address many problems successfully, in almost any field.

The second part of this chapter discussed the application of AI methods for musical purposes, which is most relevant to this study. Here, the discussions focused on three musical practices: sound synthesis, music production, and musical composition. Several works from these three categories have been reviewed to highlight the different uses of AI methods. This review has shown that learning methods and rule-based models have been the approaches most applied to address some of the challenges that can be experienced in these practices.

In summary, this background information chapter has briefly introduced the vast field of AI research and highlighted some approaches that have been utilised in music. It has shown that machine learning methods and rule-based models have been successful in addressing different musical problems, but also to identify musical characteristics. Furthermore, this chapter has highlighted some key areas that require further investigations:

- Computational representation of musical timbre's perception.
- AI methods for automatic timbre classification.
- AI methods for prediction of timbre values.
- AI methods for instrument combinations search algorithms.

4

Timbral Ranking

4.1 Chapter Overview

This chapter presents the first research development towards establishing techniques for the analysis and control of the perceptual qualities emerging from instrument timbre combinations. Firstly, the text details the need of defining methods to retrieve different timbre properties from audio files of instrument combinations. Then, it specifies the motivations for developing a computing system capable of ranking audio samples according to their timbre content. It also states the rationale for using verbal descriptors of timbre quality. The chapter continues by listing the different verbal attributes that are implemented in the timbre ranking system. Details of their respective acoustic features, with their subsequent methods of calculation, are also detailed.

The second half of the chapter presents the technical explanations of the implementation of the timbre ranking system. Here, information about the programming environment, calculation methods of timbre qualities, comparison method, and the results' ranking are provided in order to define the design and functioning of such a system. Next, the text details the methods of a perceptual experiment, designed to compare and evaluate the rankings of the system with human ratings. Results and discussions of the perceptual experiment are also provided. The chapter then concludes with discussions and reflections on the timbre ranking system and the perceptual experiment, which have informed the developments presented in the next chapter.

Below is an overview of this chapter's structure:

- 4.2 - Introduction

- 4.3 - Motivations
- 4.4 - Verbal Attributes
- 4.5 - Acoustic Features
- 4.6 - Algorithm
- 4.7 - Perceptual Experiment
- 4.8 - Chapter Conclusions
- 4.9 - Chapter Summary

4.2 Introduction

The initial step towards harnessing perceptual aspects of instrument combinations into methods for manipulating instrument timbre and timbral combinations is to determine which element of sound to utilise. This study focuses on musical timbre, which can display perceptual properties from sonic events. Therefore, it is essential to start by defining methods to retrieve, represent, and manipulate this sonic attribute.

The discussions about the notion of timbre (Section 2.4) have shown that this sonic attribute conveys important perceptual information. The review of works on musical timbre has highlighted the lack of agreed methods to represent and evaluate timbre characteristics. There are even more mixed conclusions for polyphonic timbre, which is inherent in instrument timbre combinations. Thus, this initial development investigates approaches that could be applied to evaluate polyphonic timbre from sounds of combined instruments.

The investigation has been done throughout the development of a computing system capable of ranking audio files according to specific timbre properties. This chapter details the motivations for developing such a system, the selected methods of calculation, the technical details of the timbre ranking system, and reports on a perceptual experiment designed to evaluate the approaches that have been chosen and to compare the system's results with human ratings.

4.3 Motivations

The starting point of the study presented in this thesis was the review of the few attempts at developing computing systems designed to aid orchestral compositions to understand the different approaches that have been applied to address some of the diverse orchestration challenges. It has also served to identify aspects that could benefit from incorporating elements of timbre perception of sounds from combined instruments, which motivated the implementation of the timbre ranking system presented in this chapter.

The review of works from the computer-aided orchestration field (Section 2.5) has illustrated the different approaches that have been explored since the 1970s. It was also suggested that *Orchids*[1], a system developed by IRCAM, was the most advanced computer-aided orchestration system. This program was released in November 2014 and is the result of different developments from a research project initiated in 2003. *Orchids* proposes abstract and temporal orchestrations and is also optimised for timbral mixture. In this system, the user inputs a target sound or designs an abstract target and the system tries to generate orchestrations that match that target. Different criteria help to guide the search algorithms designed to select the instrument combinations. First, *Orchids* includes a database of over 30 orchestral instruments, whose audio samples are pre-analysed and indexed by the program. This audio analysis phase provides various information, especially spectral information for each audio sample. The second set of criteria is the number and selection of the instruments that will compose the ensemble, with possible specification of pitch ranges, playing styles, and dynamics. It is also possible to spatially position the instruments, which will influence the orchestration. These different parameters guide the algorithm designed to generate instrument combinations that would match the target file. Here, the matching is based on the spectral composition of the target and of the instrument combinations. Once the search algorithm is performed, the user can listen to the solutions output by the system directly in the program, thanks to the audio samples from the database. It is also possible to render the solutions graphically into a musical score, which can be edited if required. The system also proposes to export the solutions in different file formats, such as a WAVE (Waveform Audio File Format) files and a MIDI (short for Musical Instrument Digital Interface) file, for example.

After a period of experimenting with *Orchids*, using various target sounds, shapes of abstract target, and different groups of instruments in order to observe the suggested orchestra-

[1] `http://www.forumnet.ircam.fr/product/orchids-en/`

tions, an issue started to emerge. While *Orchids* generates interesting solutions for a broad ranges of target sounds, it also proposes numerous solutions. These different orchestrations could be beneficial for inspiration if the aim is to explore various types of instrument combinations, or an unexpected instrument combination suggested by the system could be of interest. However, if the user has a specific idea in mind, it can be tedious and time-consuming to go through all the generated solutions before finding the combination that may convey the user's idea. The large number of solutions that are generated for a given target may be due to the approach selected for the search algorithm. Here, the criteria of spectrally matching the target, which is interesting and proved to perform successful results, also has a disadvantage. Several instrument and note combinations can produce similar spectral properties while they may sound very different. Thus, many solutions can fit the target file properties and be suggested by the system.

The ability to have all possible solutions should be kept because chance and surprise are part of the compositional process. However, including an option to filter the numerous solutions output by such a system would offer a way to overcome the time-consuming task of listening to a large number of solutions, which some composers may prefer to circumvent. Methods and tools developed by the Music Information Retrieval (MIR) field could provide different information to filter the solutions such as their sizes and durations for examples. However, a composer might have an idea of a sonic palette without knowing which instrument pitches to use or what duration is needed. This idea of instrument combination can be used to represent a type of sound, tone colour, or a specific perceptual quality. As discussed in Chapter 2, timbre conveys such information. Furthermore, the review of the research on timbre (Section 2.4) has shown that this sonic element is not necessarily represented by information such as duration or genre. Rather, timbre is a multidimensional element represented by combinations of different spectral and temporal information. Therefore, it could be possible to perform the known timbre calculations on the solutions output by computer-aided orchestration systems. However, composers may not necessarily be acoustician or psychoacoustician. Therefore, providing spectral and temporal information might not be an explicit filtering option. For instance, research on timbre has suggested that spectral centroid provides information about the timbre properties of a sound [100, 108], but performing its calculation and providing its resulting value may not be interpretable by many composers. Hence, it might not be very intuitive for composers to use such parameters to filter the solutions according to their sonic ideas.

From these observations, it was necessary to select an approach that alleviates the need to have expertise in acoustics or psychoacoustics to have a filter option accessible to a broad audience. In Section 2.4, it was suggested that semantics can be used to describe timbre properties. Thus, using words from the everyday language, such as brightness and roughness, could be an alternative of representing timbre properties other than with only acoustic calculations. Furthermore, these verbal descriptors also depict perceptual qualities, which could represent the sonic ideas of a composer. Instrumental mixtures can be used to convey specific textures and timbres that can display a perceptual trait. Therefore, a method that utilises perceptual qualities, represented by timbre characteristics, as criteria to filter the solutions output by computer-aided orchestration systems could be a way to overcome the time-consuming task of listening to the numerous solutions.

The following sections discuss in detail the implementation of a computing system designed to propose an approach to overcome the tedious task of listening to the numerous solutions generated by computer-aided orchestration systems. This approach utilises verbal descriptors to represent timbre properties of instrumental mixtures, instead of only using acoustic properties, and aims to alleviate the need of acoustic expertise. The different verbal descriptors that have been selected are discussed, along with descriptions of their corresponding acoustic correlate and methods of calculation. These verbal attributes are processed as parameters to filter audio files of instrument combinations. Then, technical details about the implementation of this filtering method into a computing system are provided. Finally, methods and results of a perceptual experiment conducted to evaluate the performances of the timbre ranking system are discussed.

4.4 Verbal Attributes

This section introduces the different verbal attributes, representing perceptual qualities, that have been implemented into the timbre ranking system. In Section 2.4, it was mentioned that an important amount of research has been investigating the use of semantic to represent timbre properties, which can also convey perceptual qualities [104, 107]. Research in that area has suggested several verbal descriptors, which may sometimes overlap. An extensive list of verbal descriptors, along with their correspondent acoustic correlates, can be found in [117].

With the number of verbal descriptors suggested by researchers and the potential overlap, it was necessary to make a selection. Moreover, different acoustic features have been suggested for one attribute; therefore, a preliminary implementation incorporating a few verbal attributes allows the investigation of the feasibility of this approach, before building on the advances and expanding the developments by adding further attributes. The first verbal descriptor is ***brightness***, selected due to the general agreement upon its definition and methods to represent this attribute. In regards to its use in audio, brightness can represent the presence of high frequencies in the sound and synonyms such as brilliance and clear are sometimes also used to describe this perceptual quality. The second attribute is ***dullness***, selected because it can represent qualities that are opposite to those conveyed by the term brightness, sometimes referring to a lack of brightness or quietness in the sense of not many events. Thus, this attribute is used here to contrast and compare with the attribute brightness. The third verbal descriptor is ***roughness***, a perceptual quality that has received a lot of interest in acoustics and psychoacoustics research, which suggested definitions and methods of calculation. This attribute, sometimes referring to the notion of dissonance [105, 242], can be an effect that is sought by some composers in different parts or their musical pieces. The fourth attribute is ***warmth***, a perceptual quality that is often used to describe the quality of a sound, notably appreciated by composers and musicians in some musical genres: e.g. asking sound engineers to make their music 'sound warm'. ***Warmth*** can also be a sonic texture that composers may want to include into their musical pieces. Finally, the last verbal descriptor is ***breathiness***. This attribute may relate more to sound qualities produced by voice and mouth (in singing for example). However, it also applies to some instruments, such as some brass instruments, and refers to sounds that can be produced from some of their specific playing techniques.

In total, there are five verbal attributes that have been selected: ***breathiness***, ***brightness***, ***dullness***, ***roughness***, and ***warmth***. These five attributes convey varied perceptual qualities and

offer a panel of parameters to manipulate. The next section will detail the different methods of calculation for each attribute, with discussions about the approach of estimating the timbral values from audio sources.

4.5 Acoustic Features

The previous section has introduced the different verbal attributes that have been selected and the rationale for implementing them into the timbral ranking system detailed in this chapter. This section presents their correspondent acoustic features, with explanations on their methods of calculation. The text also discusses the preferred approach for estimating the timbral values from hlaudio files.

The corresponding acoustic features and methods of calculation for each of the five timbral attributes are detailed below:

Breathiness: Research on the attribute *breathiness* suggests that the spectral shape of the sound and the combination of harmonics-to-noise ratio (HNR), calculating the amount of additive noise in a voice signal, and signal-to-noise ratio (SNR), calculating the level of background noise, provide information on the perception of this quality [243]. Here, a high amount of noise in the high frequencies, and a high amplitude between the first and second harmonic (H1 - H2) indicate a breathy sound.

Brightness: Many investigations have suggested that the spectral centroid of a sound contribute in identifying its *brightness* [100, 108]. The spectral centroid calculates the amount of high frequencies present in the sound. For a power spectrum with components $P_i(f_i)$, the spectral centroid F_c is defined as

$$F_c = \frac{\sum f_i P_i}{\sum P_i} \tag{4.1}$$

where P_i is the weighted frequency value of bin i, f_i is the center frequency of bin i, and F_c is a frequency. A high amount of high frequencies, informed by the result of the spectral centroid calculation, indicates a bright sound.

Dullness: Similar to *brightness*, the calculation of the spectral centroid of a signal provides information on estimating the *dullness* of a sound. However, in regards to this attribute, a low spectral centroid value will suggest that a sound is dull [244]. This descriptor is the inverse of the attribute *brightness*.

Roughness: The investigations on identifying the acoustic features that correspond to the perception of *roughness* suggest that the distance between adjacent partials in critical bandwidths

(the frequency bandwidth of the cochlea acting as a filter in the ear [38, 245]), which also refer to fluctuations in the signal, and sensory dissonance provide information on the *roughness* of a sound [9, 105, 246]. Here, short distances indicate the presence of a rough sound.

Warmth: Research on identifying acoustic features of the perception of *warmth* suggest the calculation of the spectral centroid of the signal and the estimation of the energy in the first three harmonics [60, 247]. Here, a low spectral centroid value, which signifies a small amount of high frequencies, and high energy in the first three harmonics indicates that a sound presents warm characteristics.

The different acoustic features for each verbal attribute have been suggested by literature investigating verbal descriptors of timbre properties and their acoustic correlates. However, as mentioned in Section 2.4, these investigations have used a broad range of sounds with diverse types of subjects, which may explain the disparity in the results and the lack of agreement in defining acoustic correlates that contribute to the perception of specific qualities. Furthermore, most of the studies that have suggested these acoustic features have investigated mostly monophonic timbre, using sets of stimuli composed of individual notes and sounds. This study aims to identify techniques to analyse and control sonic qualities of instrument combinations from processing audio files. Here, the concept of polyphonic timbre is important as multiple instruments can play simultaneously, thus creating sounds emerging from several instrument notes. However, the discussions about polyphonic timbre research (Section 2.4) have shown that there has been only a few investigations, which have suggested mixed conclusions and no consensus on the methods to estimate timbre properties from sonic events composed of a group of notes or sounds. Some investigations have suggested that methods of estimation proposed by research on monophonic timbre could be directly applied to polyphonic timbre, while others suggest performing source separation processes before applying the calculations on the individual sources. Again, these investigations have used diverse types of stimuli, which could explain the mixed conclusions. Furthermore, none of the investigations have directly investigated instrument timbre combinations, which signifies that there is no defined methods for estimating timbre characteristics for this type of sound. Thus, it was necessary to determine an approach to calculate the different acoustic properties from audio files of sounds created by combining different instruments.

For the timbral ranking system presented in this chapter, the preferred approach has been to apply the methods for calculating the acoustic features directly on the overall audio files, without performing a source separation process. First, source separation processes do not always perform accurately for every types of sound and every instruments. Furthermore, it is common to have groups of the same instrument that can play the same note, but also different notes. This combination of same instruments adds complexity to the source separation processes, which again may not be able to separate the signals. Thus, source separation is technically not a viable approach. Moreover, when listening to musical sounds, our brain may not perform a source separation to identify all individual instrumental sounds, all in a very brief amount of time. The brain may perceive and process it as a global sonic event. It can also be an artistic choice not to be able to perceive the sound as a sum of individual notes, but rather as an overall texture. Here, the notion of timbre blending is prominent, hence, estimating and summing the individual timbres may not harness the overall timbral qualities and could omit the composer's artistic choice. Finally, it is also useful to investigate if the methods suggested by monophonic timbre research could be applied directly to polyphonic timbres resulting from sounds created by instrument mixtures.

Considering the lack of standard methods, and the reasons stated in the previous paragraph, it was decided to perform the calculations of the different acoustic features on the global sounds, thus estimating the timbre properties without performing source separation. The implementation of the different acoustic features' calculations and the technical details of the timbral ranking system will be described in the following section.

4.6 Algorithm

This section proposes a technical overview of the different aspects of the timbral ranking system. First, details about the programming environment and the use of an external library are provided. Then, the text describes the different methods for estimating the timbre qualities from audio files of orchestral sounds and combinations of instrument notes. It is followed by a description of the selected approach to manipulate and compare the values resulting from the acoustic feature calculations. Finally, this section details the selected method for the representation of the results output by the system. Fig. 4.1 shows a diagram representing the flow of information of the overall algorithm designed to rank audio files according to the selected verbal attributes.

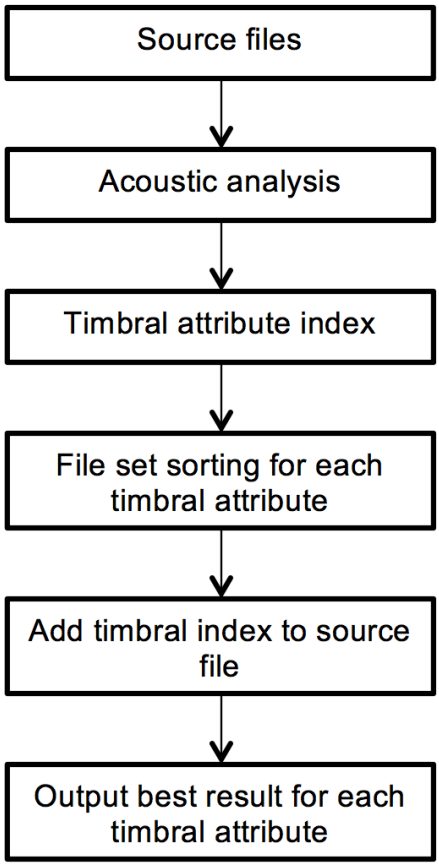

Figure 4.1: Flowchart of the system's algorithm, representing the different steps of the timbral ranking system.

4.6.1 Programming Environment

The timbral ranking system is implemented within the Matlab[1] environment on a Macintosh operating system (Mac OS). This software, which has its own programming language, has proven to be effective for numerical calculations and is widely used in diverse fields, such as engineering, for example. One of the other advantages of using this environment is the ability to add toolboxes, which can contain already made functions for specific tasks. This system uses some functions taken from the MIRtoolbox 1.6.1[2] [248], which consists of Matlab functions designed specifically for the extraction and retrieval of various musical features from audio files. The specific functions and their uses are listed in the next section. The system consists of a single script file that is executable directly in the Matlab environment. The following sections describe the different steps and functions of the system.

4.6.2 Timbre Index

This section describes the different steps of the algorithm that has been designed to retrieve the timbre qualities from audio files. Here, it has been decided to work with audio files that contain sounds created by combining instruments: audio recordings of orchestral pieces or files generated by computer-aided orchestration systems. The system processes only audio files encoded as Waveform Audio File Format (WAVE). The reason for using the WAVE format is because it is uncompressed audio, which keeps all the spectral information instead of a compressed audio format that would omit parts of the audio spectrum. For the accuracy of the acoustic feature calculations, it is essential to work with uncompressed audio, as most of the calculations are based on spectral information. The WAVE format is also an exporting option in the *Orchids* program, which motivated the development of the timbral ranking system.

The first step of the algorithm is to define the folder that contains the audio files. Here, users can select their folder via a pop-up window, which allows them to navigate through their files. Once this step is done, the system lists all the WAVE files that are present in the folder selected by the user, storing their path and name. It also creates a folder for each of the five verbal attributes. Then, for each file, the system performs the acoustic feature calculations.

The calculation processes start with using the *miraudio* function from the MIRtoolbox, which loads the audio file and stores its waveform as a variable for subsequent operations. The

[1] http://www.mathworks.com/products/matlab/

[2] MIRtoolbox is available at https://goo.gl/d61E00

reason for storing the audio file into a variable is to avoid the need to process a read and load function of the audio file for each operation. Then, the system performs the calculations of each verbal attribute incrementally. It starts with the attribute ***breathiness***, where it uses the *mirspectrum* function to obtain the audio file's spectrum, representing the distribution of the energy along the frequencies. Here, the *mirspectrum* function performs a Fast Fourier Transform (FFT). This information is then utilised to retrieve the first and second harmonic (H1 and H2) and then to calculate the amplitude between H1 and H2, and the Open Quotient (OQ) [249]. The value resulting from this calculation is then stored in a matrix variable along with the corresponding file's name. Next, the system performs the calculation for the attribute ***brightness***. Here, the system executes the *mirbrightness* function to obtain the amount of energy above a defined frequency, following suggestions by [250] and [251]. Again, the value resulting from this calculation is stored in a matrix variable along with the corresponding file's name. The same method to retrieve the amount of energy below a defined frequency is applied for the calculation of the attribute ***dullness***. For the attribute ***roughness***, the system first executes the *mirspectrum* function in order to obtain the spectrum of the audio file. Then, it performs the *mirroughness* function, which computes the average of the dissonance of all possible pairs of the spectrum's peaks [114]. Finally, the system performs the operations for the attribute ***warmth***. It starts with executing the *mirspectrum* function to obtain the audio file's spectrum. Then, it performs the *mircentroid* function onto the spectrum variable to obtain the distribution of the energy of the audio frequencies. Then, it retrieves the energy in the first three harmonics, which is subsequently combined with the value resulting from the spectral centroid calculation. Once the system has performed the verbal attributes calculations onto all the audio files contained in the selected folder, the five matrices corresponding to the five attributes are composed of all the audio files' name with their corresponding timbral value.

4.6.3 Comparison

In order to rank the audio files, the values resulting from the acoustic feature calculations need to be compared. Thus, for the matrix variable that contains the values for the attribute ***breathiness***, the matrix is indexed in descending order with highest value at the top. This value corresponds to high amount of noise in the high frequencies and high amplitude between the first and second harmonic (H1 - H2). The ***brightness*** matrix is also indexed in descending order, corresponding to the large presence of high frequencies. In regards to the attribute ***dullness***, its matrix is indexed in ascending order, which corresponds to the lack of high frequencies in the

audio file and indicates a potential dull sound. The ***roughness*** matrix is indexed in descending order, which corresponds to high fluctuations in the signal and sensory dissonance, characteristic of a rough sound. Lastly, the matrix corresponding to the attribute ***warmth*** is indexed in ascending order, corresponding to low values representing the presence of low frequencies and energy in the first three harmonics.

4.6.4 Ranking Results Output

Once the acoustic feature calculations and the indexing process have been performed, the system outputs the results of the ranking operations in two different manners.

First, for each verbal attribute, the system displays the name of the audio file that obtained the highest value. This allows the user to identify the best audio file within the several files contained in the folder. The other approach outputs the audio files with their corresponding rank number at the beginning of the name. Each attribute has its own folder. For example, in the *brightness* folder, all the audio files would be sorted from the brightest sound to the least bright sound, while the *warmth* folder will have the warmest sound at the top, and the least at the bottom. Thus, in the case of instrument combinations generated by computer-aided orchestration programs, such as *Orchids*, the user can directly listen to the solutions that most match the verbal attribute criteria, while still having the option to keep and listen to all the solutions.

4.7 Perceptual Experiment

In order to evaluate the selected methods for estimating timbre qualities, and the accuracy of the timbral ranking system, a perceptual experiment with human participants has been conducted. The purpose of this experiment was to determine the correlation between the humans' responses and the system's rankings. This section reviews this perceptual experiment, starting with details about the utilised audio stimuli. Then, the text specifies the methods that have been adopted to conduct the experiment. This section continues with presenting the results of the humans' responses. Finally, the text finishes with discussions about the experiment's results and its outcomes.

4.7.1 Training Files

It was mentioned in Section 4.3 that the development of the timbral ranking system has been motivated by experimentations with computer-aided orchestration systems, in particular with the *Orchids* program. This resulted in the generation of many audio files, some of which have been utilised for testing purposes during the implementation of the system.

For the perceptual experiment, a set of 90 audio files generated by *Orchids* has been utilised. These audio files consisted of renderings of combinations of instrument notes proposed by *Orchids*, using different targets and groups of instruments. The generated audio files were between 2 and 4 seconds long and were composed of several different instruments, such as violins, flutes, and trumpets.

The timbral ranking system has been performed on the set of 90 audio files, which resulted in having the files indexed according to their calculated timbral value for each of the five verbal attributes. Following the ranking proposed by the system, some audio files were selected as stimuli for the perceptual experiment. The choice of the audio files and the methods of the experiment will be described in the next section.

4.7.2 Methods

For each verbal attribute, three audio files were selected, following the ranking suggested by the system. Here, the file ranked number 1, number 45, and number 90 were chosen, respectively representing the highest, medium, and lowest score in each verbal attribute ranking.

The experiment were conducted with 20 individuals and performed in the same room using the same playback equipment (circumaural `Beyerdynamic DT770 PRO 250 Ohm`[1] headphones, selected for their neutral frequency responses[2], which does not colour and alter the sound). Each participant was asked to listen to 3 audio stimuli ($\approx$ 2 to 4 seconds long, with several instruments playing simultaneously) for each verbal attribute implemented in the system, which makes a total of 15 audio stimuli. For each attribute, the participants were asked to rate their respective 3 audio stimuli, using a five-point Likert scale. Fig. 4.2 shows a screenshot of the experiment's interface as viewed by participants, displaying the page for the attribute *brightness*. Each attribute's page was displayed similarly. The audio stimuli were presented in a random order for each participant, thus, avoiding a potential influence of the listening order on the participants' ratings. The results of this perceptual experiment will be presented in the next section.

4.7.3 Results

Fig. 4.3 displays the mean of the participants' ratings, using the five-point scale, for each of the three stimuli presented for the verbal attribute *breathiness*. Here, the mean rating for the stimuli *least* is 2.2, for the stimuli *medium* 2.5, and for the stimuli *most* is 3.7. Table 4.1 shows the mean and median of the participants' ratings and the standard deviation for each audio stimuli for the attribute *breathiness*.

	Perceived attribute		
Analysed attribute	Mean	Median	Standard deviation
Breathiness (least)	2.2	2	0.894
Breathiness (medium)	2.5	2	1.100
Breathiness (most)	3.7	4	0.979

Table 4.1: Mean, median, and standard deviation values of the participants' ratings for each of the three stimuli presented for the verbal attribute *breathiness*.

Fig. 4.4 displays the mean of the participants' ratings, using the five-point scale, for each of the three stimuli presented for the verbal attribute *brightness*. Here, the mean rating for the stimuli *least* is 2.4, for the stimuli *medium* 3.35, and for the stimuli *most* is 4.1. Table 4.2

[1] `http://europe.beyerdynamic.com/shop/dt-770-pro.html`
[2] `http://reference-audio-analyzer.pro/en/report/hp/beyerdynamic-dt-770-pro-250.php`

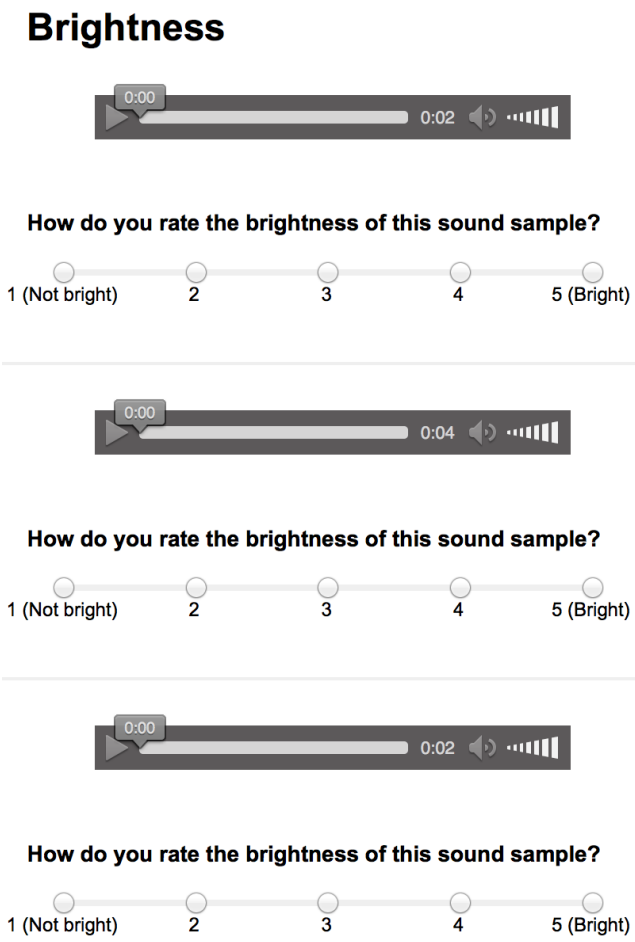

Figure 4.2: Screenshot of the experiment's interface, as viewed by participants. Here, the page for the attribute *brightness* is displayed. Pages for each attribute were displayed similarly.

shows the mean and median of the participants' ratings, and the standard deviation for each audio stimuli for the attribute *brightness*.

	Perceived attribute		
Analysed attribute	Mean	Median	Standard deviation
Brightness (least)	2.4	2	0.882
Brightness (medium)	3.35	3	0.745
Brightness (most)	4.1	4	0.641

Table 4.2: Mean, median, and standard deviation values of the participants' ratings for each of the three stimuli presented for the verbal attribute *brightness*.

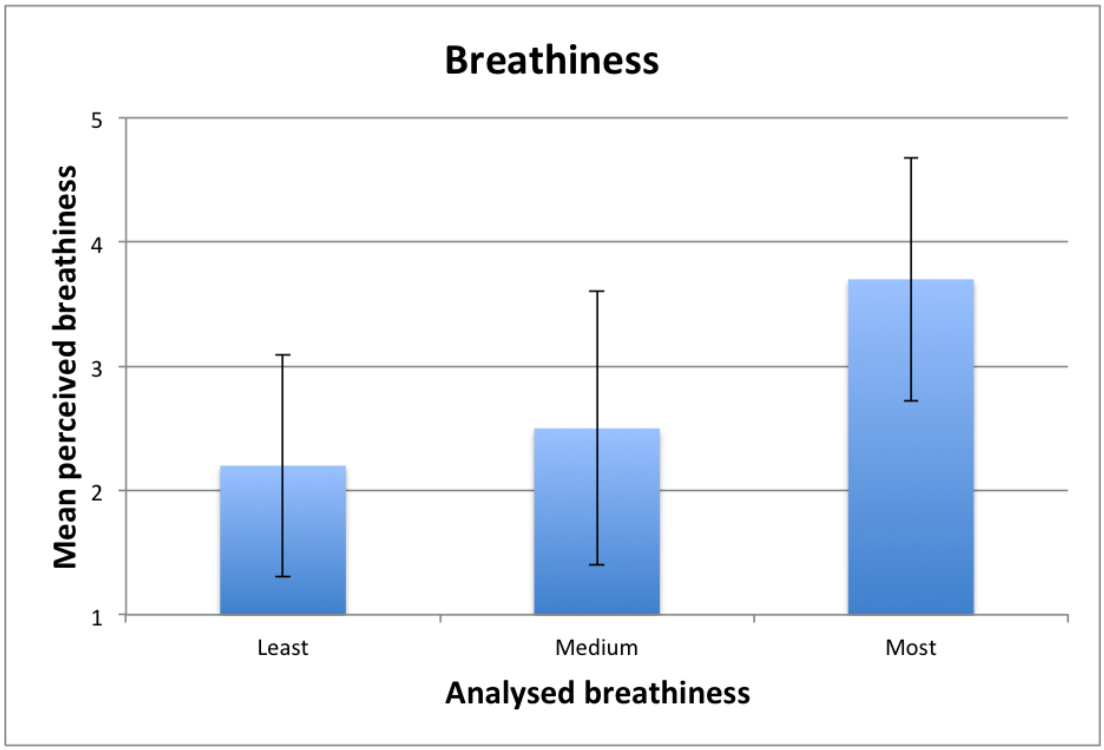

Figure 4.3: Bar graph showing the mean participants' ratings for the attribute *breathiness*. On the *x* axis are the audio stimuli as ranked by the system, and on the *y* axis is the mean of the participants' ratings, with corresponding error bars. There is a correlation between participants' ratings and the systems rankings.

Fig. 4.5 displays the mean of the participants' ratings, using the five-point scale, for each of the three stimuli presented for the verbal attribute *dullness*. Here, the mean rating for the stimuli *least* is 1.85, for the stimuli *medium* 3.2, and for the stimuli *most* is 4. Table 4.3 shows the mean and median of the participants' ratings, and the standard deviation for each audio stimuli for the attribute *dullness*.

	Perceived attribute		
Analysed attribute	Mean	Median	Standard deviation
Dullness (least)	1.85	2	1.090
Dullhtness (medium)	3.2	3	0.951
Dulltness (most)	4	4	0.645

Table 4.3: Mean, median, and standard deviation values of the participants' ratings for each of the three stimuli presented for the verbal attribute *dullness*.

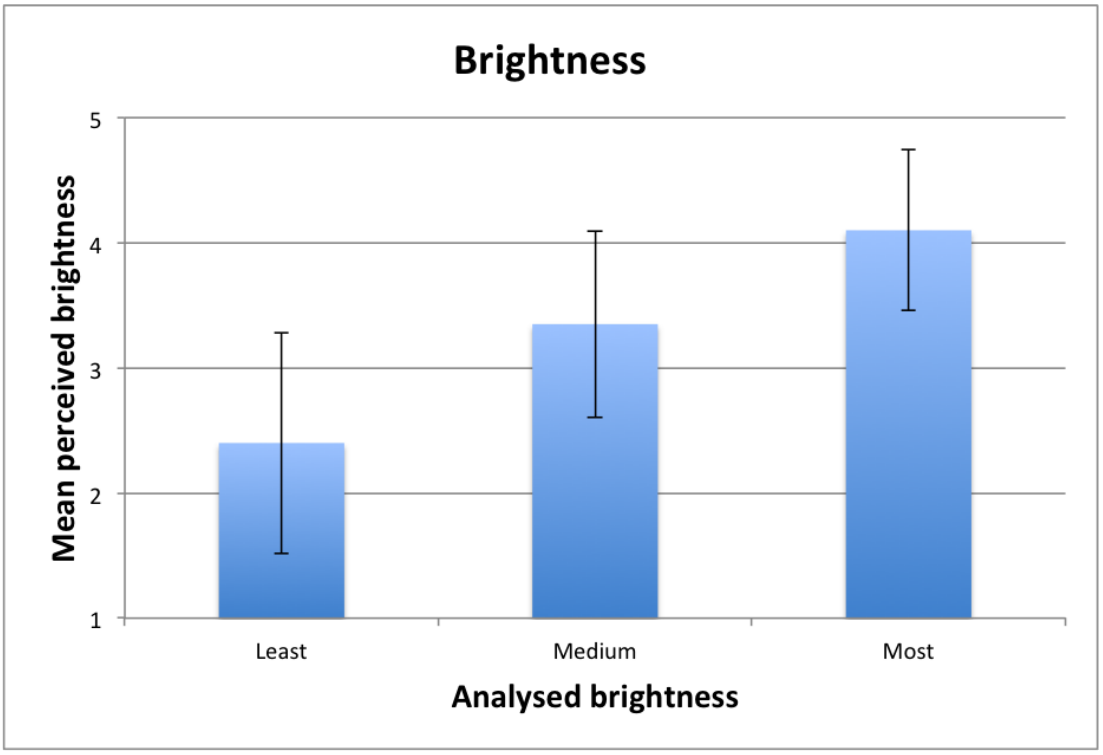

Figure 4.4: Bar graph showing the mean participants' ratings for the attribute *brightness*. On the *x* axis are the audio stimuli as ranked by the system, and on the *y* axis is the mean of the participants' ratings, with corresponding error bars. There is a strong correlation between participants' ratings and the systems rankings.

Fig. 4.6 displays the initial mean of the participants' ratings, using the five-point scale, for each of the three stimuli presented for the verbal attribute *roughness*. Here, the mean rating for the stimuli *least* is 2.25, for the stimuli *medium* is 3.7, and for the stimuli *most* is 3.5. Table 4.4 shows the mean and median of the participants' ratings, and the standard deviation for each audio stimuli for the attribute *roughness*. Following these results, the methods for estimating the attribute *roughness* have been investigated, and an error in the system's ranking has been found, which will be further detailed in the next section, along with the global discussions of the results of this perceptual experiment. Therefore, Fig. 4.7 displayed the adjusted mean of the participants' ratings. Here, the mean rating for the stimuli *least*, is still 2.25, however, for the stimuli *medium* it is now 3.5, and for the stimuli *most*, it is 3.7. Table 4.5 shows the adjusted values for the mean and median of the participants' ratings, and the standard deviation for each audio stimuli for the attribute *roughness*.

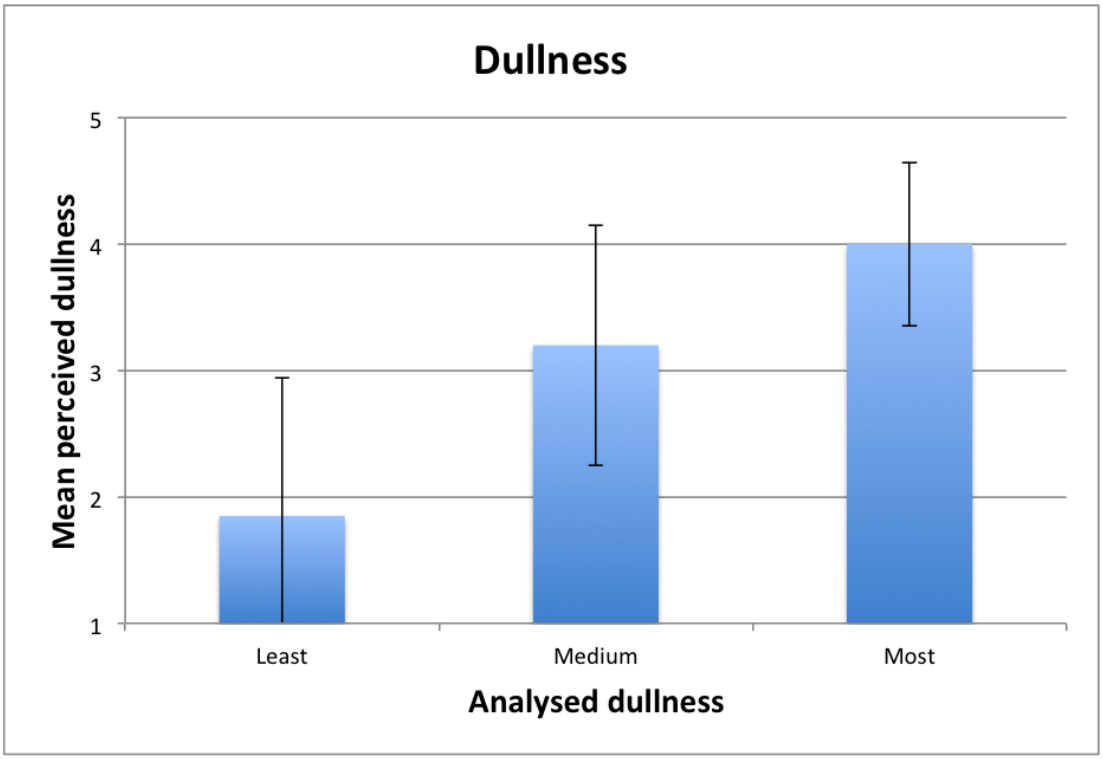

Figure 4.5: Bar graph showing the mean participants' ratings for the attribute *dullness*. On the *x* axis are the audio stimuli as ranked by the system, and on the *y* axis is the mean of the participants' ratings, with corresponding error bars. There is a strong correlation between participants' ratings and the systems rankings.

	Perceived attribute		
Analysed attribute	Mean	Median	Standard deviation
Roughness (least)	2.25	2	0.851
Roughness (medium)	3.7	4	0.946
Roughness (most)	3.5	3	1.031

Table 4.4: Initial mean, median, and standard deviation values of the participants' ratings for each of the three stimuli presented for the verbal attribute *roughness*.

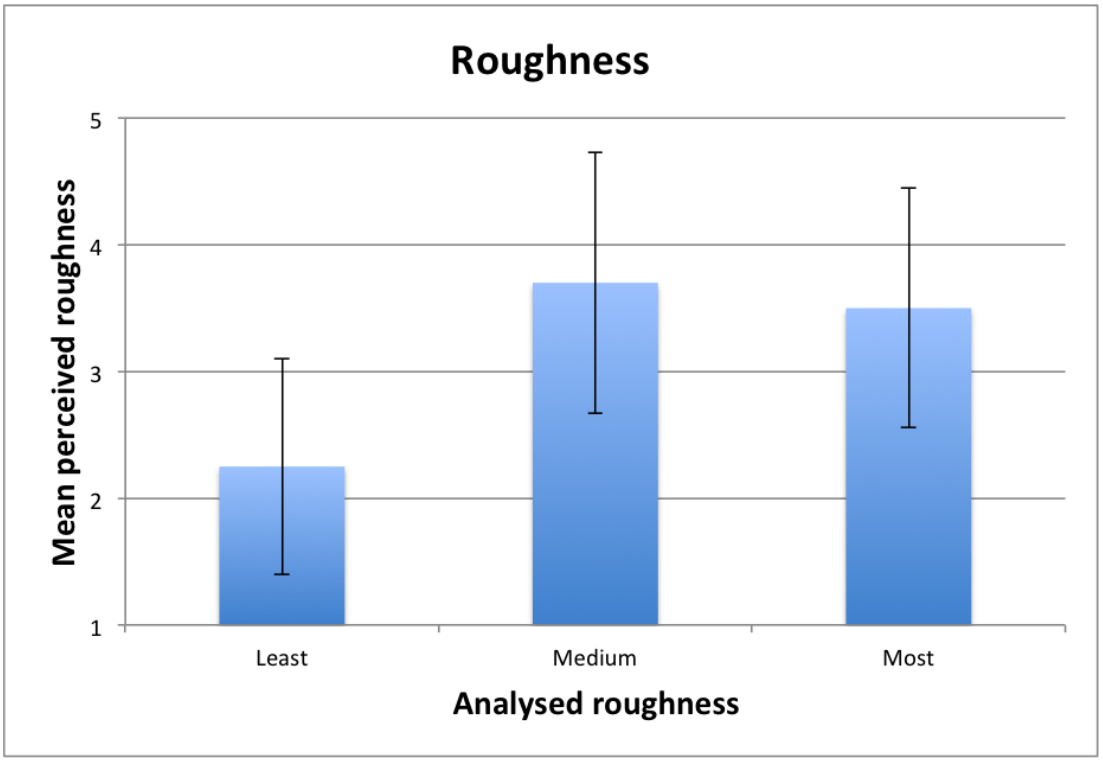

Figure 4.6: Bar graph showing the initial mean participants' ratings for the attribute *roughness*. On the *x* axis are the audio stimuli as ranked by the system, and on the *y* axis is the mean of the participants' ratings, with corresponding error bars. There is a difference between participants' ratings and the systems rankings for the audio stimuli *medium*, and *most*.

	Perceived attribute		
Analysed attribute	Mean	Median	Standard deviation
Roughness (least)	2.25	2	0.851
Roughness (medium)	3.5	3	1.031
Roughness (most)	3.7	4	0.946

Table 4.5: Revised mean, median, and standard deviation values of the participants' ratings for each of the three stimuli presented for the verbal attribute *roughness*.

Finally, Fig. 4.8 displays the mean of the participants' ratings, using the five-point scale, for each of the three stimuli presented for the verbal attribute *warmth*. Here, the mean rating for the stimuli *least* is 1.8, for the stimuli *medium* 1.8, and for the stimuli *most* is 2.75. Table 4.6 shows the mean and median of the participants' ratings, and the standard deviation for each

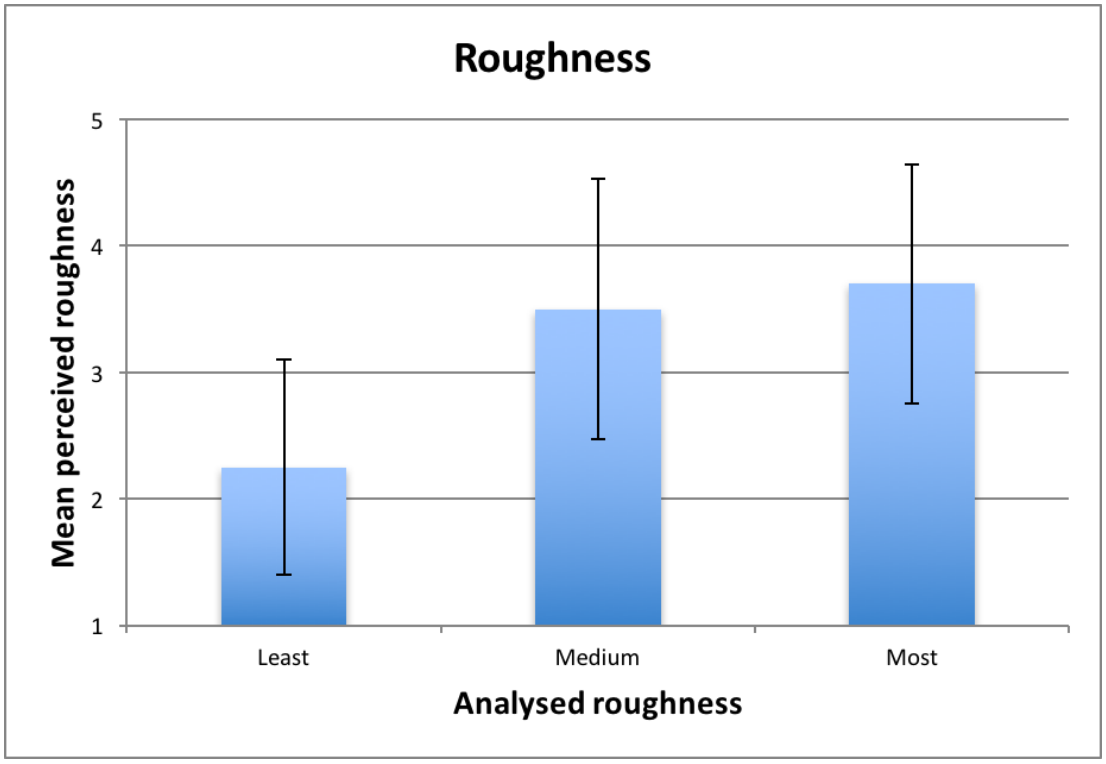

Figure 4.7: Bar graph showing the revised mean participants' ratings for the attribute *roughness*. On the *x* axis are the audio stimuli as ranked by the system, and on the *y* axis is the mean of the participants' ratings, with corresponding error bars. After adjusting the system's estimation methods, there is a correlation between participants' ratings and the systems rankings.

	Perceived attribute		
Analysed attribute	Mean	Median	Standard deviation
Warmth (least)	1.8	1	1.005
Warmth (medium)	1.8	2	0.895
Warmth (most)	2.75	2	1.165

Table 4.6: Mean, median, and standard deviation values of the participants' ratings for each of the three stimuli presented for the verbal attribute *warmth*.

audio stimuli for the attribute *warmth*.

4.7.4 Discussions

For the verbal attribute *breathiness*, the mean and median values for the audio stimuli labelled *most*, as shown in Fig. 4.3 and Table 4.1, indicate a correlation between the system's ranking

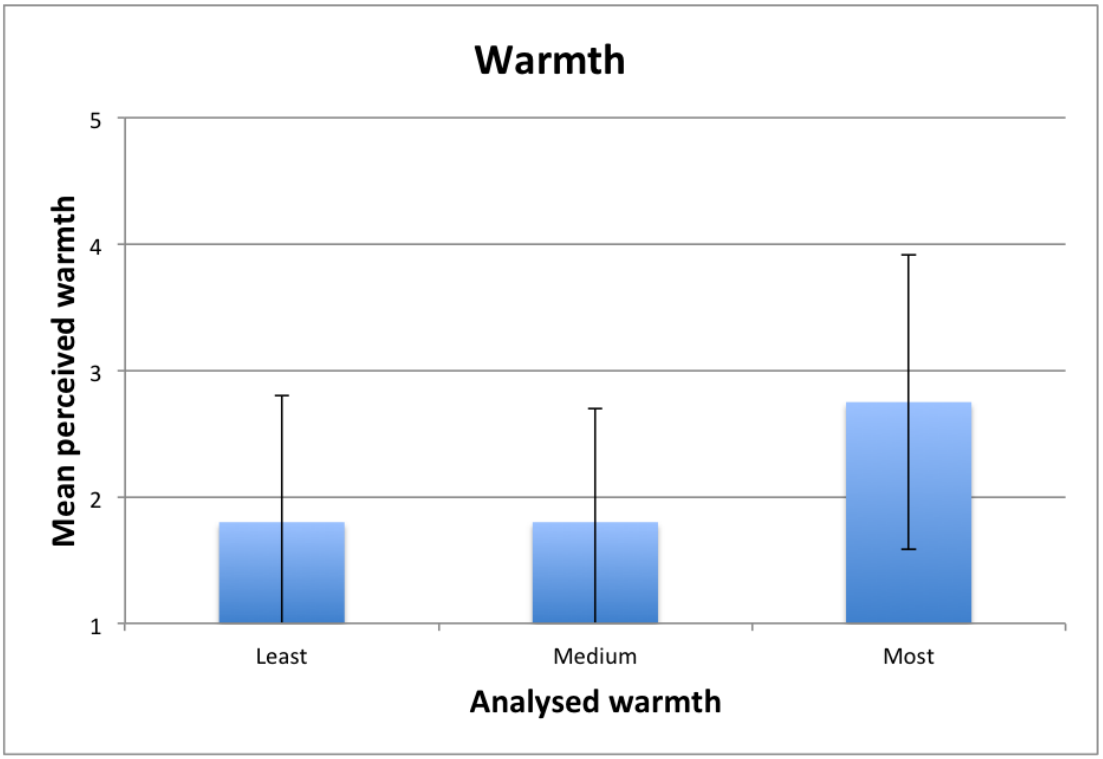

Figure 4.8: Bar graph showing the mean participants' ratings for the attribute *warmth*. On the *x* axis are the audio stimuli as ranked by the system, and on the *y* axis is the mean of the participants' ratings, with corresponding error bars. There is a difference between participants' ratings and the systems rankings for the audio stimuli *least*, and *medium*, both scored the same mean values. There is a low correlation between participants' ratings and the systems rankings for the audio stimuli *most*.

and participants' responses. Here, the file ranked as the most breathy sound by the system was also rated as breathy by the participants, with a median value of 4 out of the five-point scale. In regards to the other two audio stimuli, their corresponding median values were similar (2 out of the five-point scale), with a mean value slightly higher for the audio stimuli *medium*. The difference is too small to suggest a significant correlation between participants' responses and the system ranking. However, the similar ratings for these two stimuli could merely be due to their sonic content, which may have had a lack of breathy quality. Nevertheless, the values for the stimuli *most* suggest a correlation between participants' responses and system's rankings for the attribute *breathiness*.

Fig. 4.4 and Table 4.2 display the results for the verbal attribute *brightness*. Here, the mean values, scoring 2.4, 3.35, and 4.1 out of 5 for the stimuli *least*, *medium*, and *most* respectively,

follow the rankings proposed by the system. It is supported by the median values, where participants' responses were 2, 3, and 4 for the stimuli respectively ranked as *least*, *medium*, and *most* by the system. These results suggest a strong correlation between the participants' responses and the system's rankings for the attribute *brightness*.

The results for the verbal attribute *dullness* were similar to those for the attribute *brightness*. Here, the mean and median values, as depicted in Fig. 4.5 and Table 4.3, suggest a strong correlation between the participants' responses and the system's rankings for the attribute *dullness*.

For the verbal attribute *roughness*, the initial participants' ratings were suggesting that the system's rankings were different to the humans' responses. Fig. 4.6 and Table 4.4 display the initial values for the mean and median of the participants' ratings. Here, the mean and median values for the stimuli *least* suggested a correlation between the system' ranking and participants' responses. However, for the stimuli *medium* and *most*, there was a difference between participants' ratings and system's rankings. Here, the mean and median values for the stimuli *medium* were suggesting that actually this sound was 'rougher' than the one ranked as the most rough sound. Following these results, the methods for estimating the attribute *roughness* have been reviewed. As described in Section 4.6, the system uses the *mirroughness* function from the `MIRtoolbox` for the calculation of this attribute. It was found that the function was not returning the variable corresponding to the calculation of the average of the dissonance all possible pairs of spectrum's peaks, which was caused by specifying a wrong export parameter. Following this adjustment in the methods for estimating the attribute *roughness*, a new ranking analysis was performed on the audio stimuli, and the results of the perceptual experiment were revised. Fig. 4.7 and Table 4.5 display the mean and median values of the participants' ratings with the revised system's ranking. Here, the mean values for the stimuli *least*, *medium*, and *most* correlate with the rankings suggested by the system. This correlation is supported by the median values, where stimuli *least*, *medium*, and *most* obtained 2, 3, and 4 respectively. These revised results suggest a correlation between humans' responses and the system's rankings for the attribute *roughness*.

Finally, Fig. 4.8 and Table 4.6 display the results for the verbal attribute *warmth*. Here, the file labelled as *most* was also rated as warm by the participants, with a mean value of 2.75, and a median value of 2. In regards to the other two audio stimuli, their corresponding mean values were similar (1.8 out of the five-point scale), with a median value slightly higher for the audio stimuli *medium*. Here, the difference in the participants' ratings is too small to suggest

a significant correlation between participants' responses and the system ranking. However, the similar ratings could merely be due to the sonic content of the audio stimuli, which may have had a lack of warm properties. This is supported by the low scores for the stimuli labelled *most*. Nevertheless, considering the small differences, these values, especially for the audio stimuli *most*, suggest a small correlation between participants' responses and the system's rankings for the attribute *warmth*.

Overall, the results of the perceptual experiment suggest that the rankings proposed by the system, based on estimations of timbre characteristics, correlate with the human perception. The values of the participants' ratings for the verbal attributes *brightness* and *dullness* suggest a significant correlation between the participants' responses and the system's rankings. For the attributes *breathiness*, *roughness*, and *warmth* a correlation is also present, but with a lower significance following the different mean and median values. These results also indicate that the implemented methods for estimating timbre characteristics perform adequately with the participants' responses, suggesting that such methods could be applied to audio files generated by a computer-aided orchestration system, composed of different instruments, and notes played simultaneously.

4.8 Chapter Conclusions

The previous sections have presented the different steps of the implementation of a timbral ranking system, which serves as foundations for this study. The initial motivation for developing such a system was to overcome the task of having to go through all the numerous and diverse solutions generated by computer-aided orchestration systems, in this case with the *Orchids* program, before finding the instrument combination that corresponds to the desired type or quality of sound. The solution to address this issue was to propose a filtering option using perceptual qualities as parameters. Discussions in Section 2.4 have suggested that the sonic attribute timbre conveys perceptual information. However, it has been explained that timbre characteristics are the results of spectro-temporal properties, which are not necessarily accessible to a broad audience. Thus, in order to alleviate the need of expertise in acoustics and psychoacoustics, it was decided to use verbal descriptors of perceptual qualities, informed by timbre properties, as filtering parameters.

This study focuses on combinations of instrument notes, which consist of several instruments playing simultaneously, and due to the lack of agreed methods for calculating polyphonic timbre, it was necessary to select a few verbal attributes and test the feasibility of this approach. Thus, the methods of calculations have been selected from the literature on monophonic timbre, in order to investigate if these methods could also be applied to sounds produced by combining different instrument notes. The system integrates a few already made functions, taken from the `MIRtoolbox`, for the calculation of some attributes, as described in Section 4.6. The aim of this initial study was not to propose new methods to calculate timbre properties, but rather to see if it is possible, and how, to apply them to audio recordings of orchestral pieces and audio files of instrument note combinations.

The perceptual experiment, conducted with 20 participants, was designed to evaluate the system's rankings and the methods of timbral calculations. Its results have suggested that there is a correlation between participants' responses and the system's rankings, which indicates the methods implemented in the system can retrieve timbral properties from audio files of instrument combinations. The ratings for the verbal attributes *brightness* and *dullness* have suggested significant correlation between the participants' responses and the system's rankings. The estimating methods of these two attributes merely rely on the calculation of the spectral centroid, which has been suggested by many investigations as an acoustic correlate of these two attributes (Section 2.4). The correlation for the other attributes presented slightly

lesser significance, which may be merely due to the content of the audio stimuli. Nevertheless, the development of the timbral ranking system has provided insights for addressing **RQ1** and **RQ2** by suggesting an approach for harnessing perceptual qualities of instrument note combinations using timbre and by developing methods to compare values resulting from calculations of timbre characteristics.

One of the limitations of this timbral ranking system is the comparison of the calculated values. Here, the system only compared the timbral values resulting from the analysis of the audio files present in the folder. Thus, it can only return if a timbral value is lower or higher than another value. This approach does not provide information for estimating the dominant perceptual quality, and thus, it is not possible to classify an audio file into one category—one of the five verbal attributes. The methods presented in this chapter can retrieve timbral values. However, further investigations are required to completely harness the perception of timbre emerging from instrument combinations into techniques for the analysis and control of instrument timbre combinations.

4.9 Chapter Summary

This chapter has presented the development of a timbral ranking system, designed to investigate the possible extraction and manipulation of timbre properties from sounds emerging from instrument mixtures, which serves as the basis of the research developments presented in this thesis.

First, this chapter explained the motivations for developing such a system. The review of previous works in computer-aided orchestration and the experimentations with the *Orchids* program have shown that computer-aided orchestration systems can generate numerous solutions for a single target. The broad range of note and instrument combinations could be beneficial as a source of inspiration. However, some users may want to circumvent the tedious and time-consuming task of having to listen to all the generated solutions before finding the combination that may convey their initial idea. To overcome this issue, it has been suggested to propose a filtering option using perceptual quality criteria, which can be represented by timbre properties, as discussed in Section 2.4. In order to alleviate the need to have expertise in acoustics or psychoacoustics to utilise timbre properties, and have a filtering option accessible to a broad audience, it was decided to use verbal attributes, such as brightness, to represent timbre characteristics.

Due to the low amount of research investigating polyphonic timbre and the lack of agreed methods, it was necessary to confirm the methods of calculation for a selection of verbal attributes, before expanding the number of proposed parameters. Here, five verbal attributes have been chosen: ***breathiness***, ***brightness***, ***dullness***, ***roughness***, and ***warmth***, detailed in Section 4.4. Their selected methods of estimations have been described in Section 4.5 and the technical details in regards to the implementation of the timbral ranking system have been stated in Section 4.6.

A perceptual experiment has been conducted in order to evaluate the methods of timbre estimations, and if the rankings proposed by the system were correlated with human perception. The participants' responses, presented in Section 4.7, suggest that the selected methods for estimating timbre characteristics represented by verbal attributes can be applied to audio recordings of orchestral pieces and audio files of instrument note combinations. Furthermore, the experiment's results also suggest that the rankings proposed by the system follow human perception, which would validate the selected approach for filtering instrument combinations generated by computer-aided orchestration systems.

In summary, this chapter has described the development of a computing system designed to overcome the tedious and time-consuming task of having to listen to the numerous solutions generated by computer-aided orchestration systems. Here, verbal descriptors of perceptual qualities, represented by timbre characteristics, have been selected as filtering parameters. The results of a perceptual experiment conducted to evaluate the accuracy of the timbre estimation methods and the timbral rankings proposed by the system, have suggested that the selected approach may be suitable to address the numerous instrument note combinations listening tasks. Furthermore, the developments presented in this chapter have highlighted some key outcomes that have informed the investigations presented in Chapters 5 and 6:

- Selection of five verbal attributes to represent timbre in order to manipulate perceptual qualities.
- Methods for estimating verbal attributes, represented by timbre characteristics, from sounds produced by combining instrument notes (polyphonic timbre and timbre blending).
- Methods to address the time-consuming listening task of large sets of audio files.

5

Timbral Classification

5.1 Chapter Overview

This chapter presents an approach to automatically classify the perceptual quality of excerpts from audio recordings of orchestral pieces and audio files of instrument note combinations, represented by their timbre characteristics. Such investigation is built on the findings presented in Chapter 4 and aims to offer a solution to overcome the limitations of the initial system.

The text starts with stating the limitations that have been identified from the development of the timbral ranking system detailed in the previous chapter and defines the approach selected for developing a timbre classification. Then, the chapter presents the methods for estimating timbre characteristics from audio files, based on the findings suggested in Chapter 4. The text continues with a description of the initial methods for classifying sounds produced by combining instruments according to specific perceptual qualities, which raised some issues for data handling and did not produce successful results.

The second part of this chapter details the developments for addressing the limitations of the initial classification approach. Following discussions presented in Chapter 3, it has emerged that methods taken from the field of Artificial Intelligence (AI) research could be utilised to overcome the issues raised by the initial classification technique. Here, approaches designed to learn classification models from data have been selected. Thus, the text describes the different machine learning algorithms that have been implemented, along with discussions about their results and performances for classifying specific perceptual qualities of sounds emerging from instrument mixtures. The chapter then concludes with discussions on findings and outcomes of the development of an automatic timbre classification system, which informed the research

developments presented in the next chapter. The structure of this chapter is as follows:

- 5.2 - Introduction
- 5.3 - Timbre Estimations
- 5.4 - Initial Timbre Classification Approach
- 5.5 - Machine Learning Algorithms
- 5.6 - Chapter Conclusions
- 5.7 - Chapter Summary

5.2 Introduction

The developments presented in the previous chapter have provided methods for calculating acoustic features from audio files of instrument note combinations. The use of verbal descriptors of perceptual qualities, instead of only their acoustic features, has offered a solution to alleviate the need for acoustics and psychoacoustics expertise. Furthermore, the development of the timbral ranking system has demonstrated that the data resulting from the acoustic feature calculation could be manipulated and compared, which produced methods for establishing a ranking of numerous audio files according to their timbre content.

The initial development has provided useful insights for estimating timbre properties from audio files that consist of combination of instruments. However, it has also raised some issues that needed to be addressed. Here, the ranking was performed by comparing the timbral values from a set of audio files. While this process performed relatively well, as suggested by the results of the perceptual experiment, some limitations were evident. Due to the comparison being performed between the audio files present in a folder, if none of the files contains characteristics of a perceptual quality, the system would still output a 'top' result. For example, the system would suggest that one audio file is the brightest out of all the analysed audio files despite that the audio file may not be a bright sound. From this observation, it was essential to develop a method for identifying the dominant perceptual quality of a sound created by combining instruments, and thus be able to classify the sound into a category, according to its timbre characteristics.

The next sections present an investigation into establishing a timbre classification method for sounds emerging from instrument mixtures. The text starts with a description of the methods for estimating timbre characteristics, which were built on the outcomes of the research developments presented in Chapter 4, before continuing with details and discussions about the different timbre classification approaches that have been explored.

5.3 Timbre Estimations

This section describes the different processes designed for estimating timbre characteristics from audio files. The timbre estimations procedure is built upon the outcomes of the research developments presented in Chapter 4, and its methods are adapted from the implementation of the timbral ranking system. The text starts with details about the selected verbal attributes and their acoustic features, which have been taken from the initial research development. Then, this section provides technical details about the methods for estimating specific timbre characteristics, and thus, obtaining the timbral values.

5.3.1 Verbal Attributes

In the investigations for defining an approach to classify the perceptual quality of combined instruments sounds automatically, it has been decided to use the same verbal descriptors as those selected for the development of the timbral ranking system, detailed in Chapter 4. The rationale for this selection is that methods presented in the previous chapter have been proven to perform successfully. Furthermore, it was necessary to test the feasibility of this approach before expanding the developments by adding further attributes. Thus, the five verbal attributes that have been selected are ***breathiness***, ***brightness***, ***dullness***, ***roughness***, and ***warmth***. The next section details the methods for the acoustic feature calculations of each verbal attribute.

5.3.2 Acoustic Features

As mentioned in the previous section, there are five timbral attributes that have been selected for this investigation: *breathiness*, *brightness*, *dullness*, *roughness*, and *warmth*. The corresponding acoustic features and methods of calculation for each attribute are based on those described in Section 4.5. A brief recap is detailed below:

Breathiness: The spectral shape of the sound and the combination of harmonics-to-noise ratio (HNR) and signal-to-noise ratio (SNR) provide information on the perception of the attribute *breathiness* [243]. Here, a high amount of noise in the high frequencies and a high amplitude between the first and second harmonic (H1 - H2) indicate a breathy sound.

Brightness: The spectral centroid, which calculates the amount of high frequencies present in the sound, provides information about *brightness* [100, 108]. Here, a high amount of high

frequencies indicates a bright sound.

Dullness: Similar to *brightness*, the calculation of the spectral centroid of a signal provides information on estimating the attribute *dullness*. However, here, a low spectral centroid value will suggest that a sound is dull [244].

Roughness: The distance between adjacent partials in critical bandwidths, which also refer to fluctuations in the signal, and sensory dissonance correspond to the perception of *roughness* [9, 105, 246]. Here, short distances indicate the presence of a rough sound.

Warmth: The calculation of the spectral centroid and the energy in the first three harmonics provide data for the perception of the attribute *warmth* [60, 247]. Here, a low spectral centroid value, which signifies few high frequencies, and high energy in the first three harmonics indicates that a sound presents characteristics of the attribute *warmth*.

Following the outcomes of the developments presented in the previous chapter, the calculations of the different acoustic features are also performed on the global sounds, thus, estimating the timbre properties without performing source separation.

5.3.3 Algorithm

This section proposes a technical overview of the different aspects of the timbre estimation processes. First, the text details the programming environment. Then, the different steps designed for estimating timbre characteristics from instrument note combinations, performed to calculate the values corresponding to the five verbal attributes, are explained. Fig. 5.1 shows a diagram representing the flow of information of the algorithm designed to calculate the timbral values from audio files of instrument combinations.

5.3.3.1 Programming Environment

The timbre estimations are implemented within two programming environments, developed on a Macintosh operating system (Mac OS). The functions to calculate the acoustic features of each of the five verbal attributes are coded in the `Matlab`[1] environment, using functions

[1] `http://www.mathworks.com/products/matlab/`

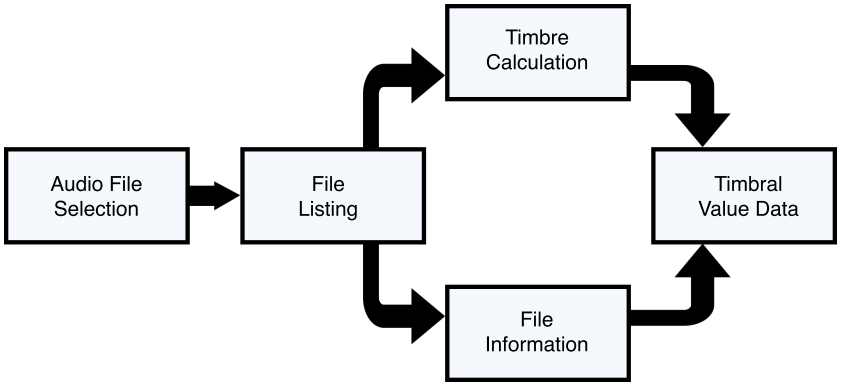

Figure 5.1: Flowchart representing the different steps to calculate timbral values from audio files.

taken from the `MIRtoolbox 1.6.1`[1], following the methods used in the timbral ranking system presented in Chapter 4. The rest of the timbre estimations functions and the whole timbral classification system are developed within the `Python 3.5` programming language. Furthermore, the `Matlab` script for the acoustic calculations is directly, and automatically, executed within the `Python` environment.

5.3.3.2 Timbral Values Estimations

This section describes the different steps of the algorithm that has been designed to estimate the timbre values, corresponding to the five verbal attributes, from sounds created by combining instrument notes. Similarly to the system presented in the previous chapter, it was decided to work with audio files that contain instrument combination sounds, whether audio recordings of orchestral pieces or files generated by computing systems. The system accepts only audio files encoded as an uncompressed WAVE audio file.

The initial step is to select the files to analyse . This step can be achieved via a pop-window, which allows the user to navigate through their files in order to select a folder containing audio files or specific audio files directly. Corresponding paths of each audio file are then stored, which are then utilised by the `Matlab` script for the acoustic feature calculations. Here, the `Matlab` processes are similar to those implemented for the timbral ranking system. The details of the specific `Matlab` process for each timbral attribute can be found in Section 4.6.2. Once the acoustic feature calculations have been performed, the `Matlab` script creates a file containing each audio file's name and the calculated timbral values for each of the five attributes. This file

[1]MIRtoolbox is available at `https://goo.gl/d61E00`

can then be further processed by functions implemented in `Python 3.5` for automatic timbre classification purposes.

5.4 Initial Timbre Classification Approach

This section presents the initial approach for automatically classifying audio files according to specific perceptual qualities that are informed by timbre characteristics. Here, the classification utilises the timbral values, calculated as described in the previous section, as input data. The following sections detail the rationale for developing such an approach, along with the description of the different methods and their results.

5.4.1 Motivations

The development of the timbral ranking system (Chapter 4) has provided methods of estimating specific timbre characteristics from audio files. It has also proposed a method for the comparison of the timbral values, which has performed successfully as suggested by the results of a perceptual experiment. However, if the audio files do not contain many characteristics of a specific perceptual quality, the system was still outputting a 'best' result. Therefore, a method for automatically classifying sounds emerging from instrument mixtures according to their perceptual qualities could be a solution to overcome this limitation.

Discussions in Chapter 2 have suggested that some acoustic features correlate with specific perceptual quality, which was confirmed by the research developments presented in the previous chapter. Thus, the timbral values obtained with the methods described in Section 5.3.2 could be used as input data for a classification method. However, with a lack of agreed metrics, which adds to the mixed conclusions about methods for estimating timbre properties as mentioned earlier, there is no known threshold to define in which category a sound could fit, based only on its calculated timbral value. Therefore, it is essential to design a method to identify the most prominent perceptual quality by analysing the timbral values of the five verbal attributes.

The following sections present the type of data that needs to be processed and the subsequent challenges. The text continues with describing the selected approach, with a review and discussions of its results.

5.4.2 Dataset

Section 5.3.2 has described the different acoustic features that correlate with the five verbal attribute selected for this investigation. These methods process and retrieve different type of information, which result in various types of data. For example, Table 5.1 presents calculated timbral values for 5 randomly selected audio files. Here, the timbral values for each verbal

	Timbral Values				
Audio File	**Breathiness**	**Brightness**	**Dullness**	**Roughness**	**Warmth**
AudioFile 1	0.101731	0.501593	0.501593	163.126015	0.009575
AudioFile 2	0.057935	0.372489	0.372489	676.719612	0.088799
AudioFile 3	0.50713	0.457623	0.457623	2641.558271	0.089005
AudioFile 4	0.086946	0.371547	0.371547	610.97954	0.123053
AudioFile 5	0.171251	0.995854	0.995854	0.002275	0.000152

Table 5.1: Examples of calculated timbral values. Note the diverse scales and figures.

attribute have diverse scales and figures, which illustrate the difficulty of establishing a classification method using only calculated timbral values.

An initial step towards establishing a scale is to identify the minimum and maximum values for each verbal attribute. Again, due to the lack of agreed metrics, these values are not defined. Therefore, it has been decided to perform an analysis of different audio files in order to gather some data and retrieve the minimum and maximum values from that dataset. Furthermore, a function has been designed to update these values each time a new audio file is analysed. This function also ensures that these metrics are more accurate with increasing data input.

Once the minimum and maximum values for each verbal attribute are defined, it is possible to rescale all the values in the range 0.0–1.0 to have similar data for each attribute. This is processed as follows:

$$\frac{x - min}{max - min} \tag{5.1}$$

where x is the calculated timbral value, *min* the known minimum value for the verbal attribute, and *max* the known maximum value for the verbal attribute. This rescaling process provides a method to manage the varied types of data, and thus, could be used to develop a method to classify the audio file. The next section describes the initial classification approach where distance calculations have been used to identify the dominant perceptual quality.

5.4.3 Distance Calculations

One method to identify the dominant perceptual quality is to measure each attribute's distance from the calculated timbral value with the known maximum value of the attribute, which would determine how close the timbral value is with the highest known value. Thus, the attribute with the shortest distance would be classified as the dominant quality as it is the closest to

the 'ideal value'. The following sections detail the different distance functions that have been implemented. The results of their testing are also presented.

5.4.3.1 Euclidean Distance

The first distance function that has been implemented is the Euclidean distance, which returns the straight-line distance between two points in a plane space. Here, the distance between the timbral value and the known maximum value is calculated as follows:

$$d = \sqrt{\sum (x_i - y_i)^2} \tag{5.2}$$

where x_i is the known maximum value for the verbal attribute and y_i is the calculated timbral value. The distance function is applied to each timbral value of each verbal attribute. Then, the function returns the verbal attribute that has the shortest distance to their known maximum timbral value, and thus, classify the audio files according to their dominant perceptual quality.

In order to test the distance calculations as a method of classification, 50 audio recordings of diverse orchestral pieces have been selected in order to represent instrument timbre and timbral combinations. Here, the audio recordings have been split into 1, 2, 3, 4, and 5 second audio files, adding a 50 ms fade in and fade out for each file. The rationale for splitting the recordings is due to the acoustic features properties. The analysis of long audio sources would not provide accurate values as some acoustic features are time-related. Therefore, it is essential to split them into shorter audio samples. Furthermore, the different lengths have been chosen to match the duration of outputs generated by computer-aided orchestration systems, but also to determine if the difference in time duration would have an impact on the results. These audio splitting can also be used to examine the variations of perceptual qualities throughout the compositions.

Table 5.2 lists the results of the classification process performed on 4-second split audio files. Here, using the method described in Section 5.4.2, the distance calculations have been applied on rescaled timbral values. For each orchestral piece, the table shows the number of audio files that have been classified in each verbal attribute category. The highest number (in red) corresponds to the most dominant perceptual quality throughout the piece.

Similar testing of distance calculations, using the same set of audio files, have been applied on unscaled data in order to determine if the rescaling process was altering the data. Table 5.3 lists the results of the classification process performed on 4-second split audio files, using

Composition	Breathiness	Brightness	Dullness	Roughness	Warmth
Orff: Carmina Burana - O Fortruna	0	5	**36**	0	0
Debussy: Suite Bergamasque - Clair De Lune	0	1	**78**	0	0
Verdi: Nabucco - Chorus Of The Hebrew Slaves	0	1	**74**	0	0
Holst: The Planets - Jupiter	0	11	**106**	0	0
Dvorak: Symphony #9, "From The New World" - 2. Largo	1	4	**188**	0	0
Bach: Orchestral Suite #3 - Air	0	1	**63**	0	0
Grieg: Peer Gynt - Morning Mood	0	1	**55**	0	0
Beethoven: Symphony #5 - Allegro Con Brio	0	1	**111**	0	0
Chopin: Nocturne #2	0	3	**71**	0	0
Pachelbel: Canon	0	3	**50**	0	0
Barber: Adagio For Strings	1	24	**101**	0	0
Vivaldi: Violin Concerto In E, Op. 8:1, RV 269, "The Four Seasons (Spring)" - 1. Allegro	1	12	**35**	0	0
Wagner: Ride Of The Valkyries	0	26	**52**	0	0
Bach: Brandenburg Concerto #3 - Allegro	0	0	**86**	0	0
Tchaikovsky: Swan Lake Suite - Scene	0	1	**45**	0	0
Satie: Gymnopedie #1	0	0	**55**	0	0
Beethoven: Symphony #9 In D, "Choral" - Finale: Ode To Joy	0	13	**25**	0	0
Mozart: Piano Concerto #21 - 2. Andante	0	1	**109**	0	0
Brahms: Hungarian Dance #5	0	2	**38**	0	0
Massenet: Thais - Meditation	2	13	**61**	0	0
Elgar: Pomp & Circumstance March #1	0	31	**56**	0	0
Mozart: Requiem Mass - Lacrimosa	0	1	**42**	0	0
Strauss Jr. (J): The Beautiful Blue Danube	0	9	**79**	0	0
Beethoven: Fur Elise	0	1	**48**	0	0
Bizet: Carmen Suite - Habanera	0	6	**27**	0	0
Rossini: The Barber of Seville: Overture	1	15	**89**	0	0
Myers: Cavantina	0	0	**60**	0	0
Dvorak: Slavoni Dance No. 2	0	0	**84**	0	0
Verdi: Medda de Requiem	0	3	**53**	3	0
Mahler: Symphony No. 5	0	11	**152**	0	0
Gounod: Ave Maria	0	0	**39**	0	0
Grieg: In the Hall of the Mountain King	3	7	**31**	0	0
Beethoven: Moonlight Sonata: Adagio Sostenuto	0	0	**84**	0	0
Mozart: Eine Kleine Nachtmusik: Allegro	0	4	**84**	0	0
Giazotto: Adagio in G Minor,	2	8	**118**	0	0
Sibelius: Finlandia	0	11	**110**	0	0
Boccherini: String Quartet: Minuet	0	1	**56**	0	0
Mozart: Symphony No. 40	0	1	**105**	0	0
Smetana: Ma Vlast - Vltava	0	13	**159**	0	0
Rachmaninov: Vocalise	0	8	**80**	0	0
Beethoven: Egmont Overture	0	4	**116**	0	0
Handel: Messiah - Hallelujah Chorus	0	1	**52**	0	0
Faure: Pavane	0	1	**76**	0	0
Bach: Concerto For 2 Violins - Vivace	0	2	**53**	0	0
Offenbach: The Tales Of Hoffman - Barcarolle	0	3	**51**	0	0
Mozart: The Magic Flute - Overture	0	1	**95**	0	0
Mozart: Piano Sonata #11 - 3. Rondo Alla Turca	0	1	**51**	0	0
Bizet: L'Arlesienne Suite#1 - Prelude	1	2	**101**	0	0
Corelli: Christmas Concerto - Allegro	0	6	**31**	0	0
Strauss Sr. (J): Radetzky March	0	23	**24**	0	0

Table 5.2: The results of the Euclidean distance-based classification process performed on 4-second split audio files. It displays the numbers of audio samples classified in each attribute. The most dominant attribute for each piece is highlighted in red.

unscaled timbral values. Results are presented in a similar layout as the results for the rescaled timbral values shown in Table 5.2.

From the results displayed in Table 5.2 and 5.3, it is apparent that such distance calculations did not perform successfully as all the orchestral pieces were dominantly *dull*. Therefore, a different distance function needed to be implemented in order to determine if the euclidean method is not adapted to the type of data represented by the timbral values.

5.4.3.2 Sum of Squared Difference (SSD)

The previous method of classification has utilised Euclidean distance calculations, which did not perform successfully. To determine if the issue is the Euclidean method, which in this case considers timbral values as points in plane space, a different measure derived from the Euclidean distance method has been implemented: Sum of Squared Difference (SSD). Here, squaring the values emphasises larger differences, which is not accomplished by the Euclidean distance. The SSD is calculated as follows:

$$d = \sum (x_i - y_i)^2 \tag{5.3}$$

where x_i is the known maximum value for the verbal attribute and y_i is the calculated timbral value. The measure is performed on each timbral value of the five verbal attributes, and the function returns the verbal attribute with the lowest SSD value, which represents the dominant perceptual quality.

Similarly to the euclidean distance, testing of the SSD measure was carried out using an identical method with the same 50 audio recordings and splitting options. Table 5.4 displays the results of the classification process performed on 4 second split audio files for rescaled timbral values, and Table 5.5 shows the results for unscaled timbral values. Similar to the Euclidean distance method, the testing results suggest that SDD is not a valid approach for classification based on timbral values, and measures based on euclidean distance are not adapted for timbral values data.

5.4.3.3 Sum of Absolute Difference (SAD)

Following the results of the previous classification methods using the Euclidean distance calculation and the SSD method, a different distance function has been implemented: Sum of Absolute Difference (SAD). This measure is usually utilised in digital image processing in

Composition	Breathiness	Brightness	Dullness	Roughness	Warmth
Orff: Carmina Burana - O Fortruna	0	5	**36**	0	0
Debussy: Suite Bergamasque - Clair De Lune	0	1	**78**	0	0
Verdi: Nabucco - Chorus Of The Hebrew Slaves	0	1	**74**	0	0
Holst: The Planets - Jupiter	0	14	**103**	0	0
Dvorak: Symphony #9, "From The New World" - 2. Largo	1	5	**187**	0	0
Bach: Orchestral Suite #3 - Air	0	2	**62**	0	0
Grieg: Peer Gynt - Morning Mood	0	1	**55**	0	0
Beethoven: Symphony #5 - Allegro Con Brio	0	1	**111**	0	0
Chopin: Nocturne #2	0	3	**71**	0	0
Pachelbel: Canon	0	3	**50**	0	0
Barber: Adagio For Strings	1	25	**100**	0	0
Vivaldi: Violin Concerto In E, Op. 8:1, RV 269, "The Four Seasons (Spring)" - 1. Allegro	1	14	**33**	0	0
Wagner: Ride Of The Valkyries	0	29	**49**	0	0
Bach: Brandenburg Concerto #3 - Allegro	0	0	**86**	0	0
Tchaikovsky: Swan Lake Suite - Scene	0	1	**45**	0	0
Satie: Gymnopedie #1	0	0	**55**	0	0
Beethoven: Symphony #9 In D, "Choral" - Finale: Ode To Joy	0	16	**22**	0	0
Mozart: Piano Concerto #21 - 2. Andante	0	1	**109**	0	0
Brahms: Hungarian Dance #5	0	2	**38**	0	0
Massenet: Thais - Meditation	2	13	**61**	0	0
Elgar: Pomp & Circumstance March #1	0	35	**52**	0	0
Mozart: Requiem Mass - Lacrimosa	0	1	**42**	0	0
Strauss Jr. (J): The Beautiful Blue Danube	0	9	**79**	0	0
Beethoven: Fur Elise	0	1	**48**	0	0
Bizet: Carmen Suite - Habanera	0	6	**27**	0	0
Rossini: The Barber of Seville: Overture	1	16	**88**	0	0
Myers: Cavantina	0	0	**60**	0	0
Dvorak: Slavoni Dance No. 2	0	0	**84**	0	0
Verdi: Medda de Requiem	0	3	**52**	4	0
Mahler: Symphony No. 5	0	14	**149**	0	0
Gounod: Ave Maria	0	0	**39**	0	0
Grieg: In the Hall of the Mountain King	3	9	**29**	0	0
Beethoven: Moonlight Sonata: Adagio Sostenuto	0	0	**84**	0	0
Mozart: Eine Kleine Nachtmusik: Allegro	0	6	**82**	0	0
Giazotto: Adagio in G Minor,	2	12	**114**	0	0
Sibelius: Finlandia	0	12	**109**	0	0
Boccherini: String Quartet: Minuet	0	1	**56**	0	0
Mozart: Symphony No. 40	0	1	**105**	0	0
Smetana: Ma Vlast - Vltava	0	14	**158**	0	0
Rachmaninov: Vocalise	0	8	**80**	0	0
Beethoven: Egmont Overture	0	6	**114**	0	0
Handel: Messiah - Hallelujah Chorus	0	1	**52**	0	0
Faure: Pavane	0	2	**75**	0	0
Bach: Concerto For 2 Violins - Vivace	0	3	**52**	0	0
Offenbach: The Tales Of Hoffman - Barcarolle	0	3	**51**	0	0
Mozart: The Magic Flute - Overture	0	1	**95**	0	0
Mozart: Piano Sonata #11 - 3. Rondo Alla Turca	0	1	**51**	0	0
Bizet: L'Arlesienne Suite#1 - Prelude	1	3	**100**	0	0
Corelli: Christmas Concerto - Allegro	0	8	**29**	0	0
Strauss Sr. (J): Radetzky March	0	23	**24**	0	0

Table 5.3: The results of the Euclidean distance-based classification process performed on 4-second split audio files, using unscaled data. It displays the numbers of audio samples classified in each attribute. The most dominant attribute for each piece is highlighted in red.

Composition	Breathiness	Brightness	Dullness	Roughness	Warmth
Bach: Brandenburg Concerto #3 - Allegro	1	0	**85**	0	0
Orff: Carmina Burana - O Fortruna	0	4	**34**	3	0
Rossini: The Barber of Seville: Overture	1	15	**89**	0	0
Smetana: Má Vlast - Vltava	0	13	**159**	0	0
Bach: Orchestral Suite #3 - Air	0	1	**63**	0	0
Fauré: Pavane	1	1	**75**	0	0
Gounod: Ave Maria	1	0	**38**	0	0
Mozart: Piano Concerto #21 - 2. Andante	0	1	**109**	0	0
Bach: Concerto For 2 Violins - Vivace	0	2	**53**	0	0
Brahms: Hungarian Dance #5	0	2	**38**	0	0
Grieg: In the Hall of the Mountain King	3	7	**31**	0	0
Grieg: Peer Gynt - Morning Mood	1	0	**55**	0	0
Beethoven: Moonlight Sonata: Adagio Sostenuto	3	0	**81**	0	0
Beethoven: Symphony #5 - Allegro Con Brio	0	1	**111**	0	0
Massenet: Thais - Meditation	3	14	**59**	0	0
Offenbach: The Tales Of Hoffman - Barcarolle	0	3	**51**	0	0
Chopin: Nocturne #2	1	3	**70**	0	0
Elgar: Pomp & Circumstance March #1	1	31	**55**	0	0
Mozart: Eine Kleine Nachtmusik: Allegro	0	4	**84**	0	0
Mozart: The Magic Flute - Overture	0	1	**95**	0	0
Giazotto: Adagio in G Minor	8	8	**112**	0	0
Mozart: Piano Sonata #11 - 3. Rondo Alla Turca	0	1	**51**	0	0
Mozart: Requiem Mass - Lacrimosa	0	1	**42**	0	0
Pachelbel: Canon	1	3	**49**	0	0
Barber: Adagio For Strings	4	24	**98**	0	0
Bizet: L'Arlésienne Suite #1 - Prelude	1	2	**101**	0	0
Sibelius: Finlandia	0	11	**110**	0	0
Strauss Jr. (J): The Beautiful Blue Danube	0	9	**79**	0	0
Beethoven: Fur Elise	0	1	**48**	0	0
Boccherini: String Quartet: Minuet	0	1	**56**	0	0
Corelli: Christmas Concerto - Allegro	0	6	**31**	0	0
Vivaldi: Violin Concerto In E, Op. 8:1, RV 269, "The Four Seasons (Spring)" - 1. Allegro	1	12	**35**	0	0
Bizet: Carmen Suite - Habanera	0	6	**27**	0	0
Mozart: Symphony No. 40	0	1	**105**	0	0
Strauss Sr. (J): Radetzky March	0	23	**24**	0	0
Wagner: Ride Of The Valkyries	0	26	**52**	0	0
Debussy: Suite Bergamasque - Clair De Lune	1	1	**77**	0	0
Myers: Cavantina	0	0	**60**	0	0
Rachmaninov: Vocalise	1	8	**79**	0	0
Tchaikovsky: Swan Lake Suite - Scene	0	1	**45**	0	0
Beethoven: Egmont Overture	3	4	**113**	0	0
Dvořák: Slavoni Dance No. 2	0	0	**84**	0	0
Satie: Gymnopedie #1	0	0	**55**	0	0
Verdi: Nabucco - Chorus Of The Hebrew Slaves	0	1	**74**	0	0
Beethoven: Symphony #9 In D, "Choral" - Finale: Ode To Joy	0	13	**25**	0	0
Handel: Messiah - Hallelujah Chorus	0	1	**52**	0	0
Holst: The Planets - Jupiter	0	11	**106**	0	0
Verdi: Medda de Requiem	0	3	**42**	14	0
Dvořák: Symphony #9, "From The New World" - 2. Largo	8	4	**180**	1	0
Mahler: Symphony No. 5	8	11	**144**	0	0

Table 5.4: The results of the Sum of Squared Difference distance-based classification process performed on 4-second split audio files, using scaled data.

Composition	Breathiness	Brightness	Dullness	Roughness	Warmth
Bach: Brandenburg Concerto #3 - Allegro	1	0	**85**	0	0
Orff: Carmina Burana - O Fortruna	0	4	**34**	3	0
Rossini: The Barber of Seville: Overture	1	15	**89**	0	0
Smetana: Má Vlast - Vltava	0	13	**159**	0	0
Bach: Orchestral Suite #3 - Air	0	1	**63**	0	0
Fauré: Pavane	1	1	**75**	0	0
Gounod: Ave Maria	1	0	**38**	0	0
Mozart: Piano Concerto #21 - 2. Andante	0	1	**109**	0	0
Bach: Concerto For 2 Violins - Vivace	0	2	**53**	0	0
Brahms: Hungarian Dance #5	0	2	**38**	0	0
Grieg: In the Hall of the Mountain King	3	7	**31**	0	0
Grieg: Peer Gynt - Morning Mood	1	0	**55**	0	0
Beethoven: Moonlight Sonata: Adagio Sostenuto	3	0	**81**	0	0
Beethoven: Symphony #5 - Allegro Con Brio	0	1	**111**	0	0
Massenet: Thais - Meditation	3	14	**59**	0	0
Offenbach: The Tales Of Hoffman - Barcarolle	0	3	**51**	0	0
Chopin: Nocturne #2	1	3	**70**	0	0
Elgar: Pomp & Circumstance March #1	1	31	**55**	0	0
Mozart: Eine Kleine Nachtmusik: Allegro	0	4	**84**	0	0
Mozart: The Magic Flute - Overture	0	1	**95**	0	0
Giazotto: Adagio in G Minor	8	8	**112**	0	0
Mozart: Piano Sonata #11 - 3. Rondo Alla Turca	0	1	**51**	0	0
Mozart: Requiem Mass - Lacrimosa	0	1	**42**	0	0
Pachelbel: Canon	1	3	**49**	0	0
Barber: Adagio For Strings	4	24	**98**	0	0
Bizet: L'Arlésienne Suite #1 - Prelude	1	2	**101**	0	0
Sibelius: Finlandia	0	11	**110**	0	0
Strauss Jr. (J): The Beautiful Blue Danube	0	9	**79**	0	0
Beethoven: Fur Elise	0	1	**48**	0	0
Boccherini: String Quartet: Minuet	0	1	**56**	0	0
Corelli: Christmas Concerto - Allegro	0	6	**31**	0	0
Vivaldi: Violin Concerto In E, Op. 8:1, RV 269, "The Four Seasons (Spring)" - 1. Allegro	1	12	**35**	0	0
Bizet: Carmen Suite - Habanera	0	6	**27**	0	0
Mozart: Symphony No. 40	0	1	**105**	0	0
Strauss Sr. (J): Radetzky March	0	23	**24**	0	0
Wagner: Ride Of The Valkyries	0	26	**52**	0	0
Debussy: Suite Bergamasque - Clair De Lune	1	1	**77**	0	0
Myers: Cavantina	0	0	**60**	0	0
Rachmaninov: Vocalise	1	8	**79**	0	0
Tchaikovsky: Swan Lake Suite - Scene	0	1	**45**	0	0
Beethoven: Egmont Overture	3	4	**113**	0	0
Dvořák: Slavoni Dance No. 2	0	0	**84**	0	0
Satie: Gymnopedie #1	0	0	**55**	0	0
Verdi: Nabucco - Chorus Of The Hebrew Slaves	0	1	**74**	0	0
Beethoven: Symphony #9 In D, "Choral" - Finale: Ode To Joy	0	13	**25**	0	0
Handel: Messiah - Hallelujah Chorus	0	1	**52**	0	0
Holst: The Planets - Jupiter	0	11	**106**	0	0
Verdi: Medda de Requiem	0	3	**42**	14	0
Dvořák: Symphony #9, "From The New World" - 2. Largo	8	4	**180**	1	0
Mahler: Symphony No. 5	8	11	**144**	0	0

Table 5.5: The results of the Sum of Squared Difference distance-based based classification process performed on 4-second split audio files, using unscaled data.

order to identify the similarity between blocks. For example, this measure is used for object recognition. Thus, timbral values would be considered as a block of data, not points in a plane space. The SAD is calculated as follows:

$$d = \sum |x_i - y_i| \tag{5.4}$$

where x_i is the known maximum value for the verbal attribute and y_i is the calculated timbral value. The measure function is applied to each timbral value of each verbal attribute. Then, the function returns the attribute with the lowest SAD value as the dominant perceptual quality.

Similarly to the previous two methods, some testing has been performed for the SAD measure using the same 50 audio recordings and splitting options. Table 5.6 displays the results of the classification process performed on the 4 second split audio files for rescaled timbral values, and Table 5.7 shows the results for unscaled timbral values. Similar to the Euclidean distance and SSD, the testing results presented in Tables 5.6 and 5.7 suggest that the dominant perceptual quality throughout all the orchestral pieces is *dullness*, which indicates that SAD is not a valid approach for classification based on timbral values.

5.4.4 Discussions

The initial timbre classification approach utilised distance metrics as a function to identify the dominant perceptual quality from sounds emerging from instrument combinations. With the known maximum value for each of the five verbal attributes and a method to rescale all the calculated values, it is possible to measure the distance that separate the calculated timbral value with its known maximum value. These distance metrics would indicate how close the audio file is to the 'best result' represented by the known maximum timbral value. Therefore, by comparing the distance metrics, the attribute with the shortest distance would be classified as the dominant perceptual quality.

The first distance calculation that has been implemented was the Euclidean distance. The results of its testing, performed on a set of 50 audio recordings of diverse orchestral pieces using different durations for the splitting option, have suggested that the dominant perceptual quality throughout all the musical pieces was *dullness*. From this observation, it is evident that Euclidean distance is not adapted for classification using rescaled timbral values. In order to determine if the rescaling method was altering the data, similar testing has been performed using unscaled timbral values. Again, the suggested dominant perceptual quality of all the musical pieces was *dullness*. Following these results, two other distance functions have been

Composition	Breathiness	Brightness	Dullness	Roughness	Warmth
Bach: Brandenburg Concerto #3 - Allegro	1	0	**85**	0	0
Orff: Carmina Burana - O Fortruna	0	4	**34**	3	0
Rossini: The Barber of Seville: Overture	1	15	**89**	0	0
Smetana: Má Vlast - Vltava	0	13	**159**	0	0
Bach: Orchestral Suite #3 - Air	0	1	**63**	0	0
Fauré: Pavane	1	1	**75**	0	0
Gounod: Ave Maria	1	0	**38**	0	0
Mozart: Piano Concerto #21 - 2. Andante	0	1	**109**	0	0
Bach: Concerto For 2 Violins - Vivace	0	2	**53**	0	0
Brahms: Hungarian Dance #5	0	2	**38**	0	0
Grieg: In the Hall of the Mountain King	3	7	**31**	0	0
Grieg: Peer Gynt - Morning Mood	1	0	**55**	0	0
Beethoven: Moonlight Sonata: Adagio Sostenuto	3	0	**81**	0	0
Beethoven: Symphony #5 - Allegro Con Brio	0	1	**111**	0	0
Massenet: Thais - Meditation	3	14	**59**	0	0
Offenbach: The Tales Of Hoffman - Barcarolle	1	3	**50**	0	0
Chopin: Nocturne #2	1	3	**70**	0	0
Elgar: Pomp & Circumstance March #1	1	31	**55**	0	0
Mozart: Eine Kleine Nachtmusik: Allegro	0	4	**84**	0	0
Mozart: The Magic Flute - Overture	0	1	**95**	0	0
Giazotto: Adagio in G Minor	8	8	**112**	0	0
Mozart: Piano Sonata #11 - 3. Rondo Alla Turca	0	1	**51**	0	0
Mozart: Requiem Mass - Lacrimosa	0	1	**42**	0	0
Pachelbel: Canon	1	3	**49**	0	0
Barber: Adagio For Strings	4	24	**98**	0	0
Bizet: L'Arlésienne Suite #1 - Prelude	1	2	**101**	0	0
Sibelius: Finlandia	0	11	**110**	0	0
Strauss Jr. (J): The Beautiful Blue Danube	0	9	**79**	0	0
Beethoven: Fur Elise	0	1	**48**	0	0
Boccherini: String Quartet: Minuet	0	1	**56**	0	0
Corelli: Christmas Concerto - Allegro	0	6	**31**	0	0
Vivaldi: Violin Concerto In E, Op. 8:1, RV 269, "The Four Seasons (Spring)" - 1. Allegro	1	12	**35**	0	0
Bizet: Carmen Suite - Habanera	0	6	**27**	0	0
Mozart: Symphony No. 40	0	1	**105**	0	0
Strauss Sr. (J): Radetzky March	0	23	**24**	0	0
Wagner: Ride Of The Valkyries	0	26	**52**	0	0
Debussy: Suite Bergamasque - Clair De Lune	1	1	**77**	0	0
Myers: Cavantina	0	0	**60**	0	0
Rachmaninov: Vocalise	1	8	**79**	0	0
Tchaikovsky: Swan Lake Suite - Scene	0	1	**45**	0	0
Beethoven: Egmont Overture	3	4	**113**	0	0
Dvořák: Slavoni Dance No. 2	0	0	**84**	0	0
Satie: Gymnopedie #1	0	0	**55**	0	0
Verdi: Nabucco - Chorus Of The Hebrew Slaves	0	1	**74**	0	0
Beethoven: Symphony #9 In D, "Choral" - Finale: Ode To Joy	0	13	**25**	0	0
Handel: Messiah - Hallelujah Chorus	0	1	**52**	0	0
Holst: The Planets - Jupiter	0	11	**106**	0	0
Verdi: Medda de Requiem	0	3	**42**	14	0
Dvořák: Symphony #9, "From The New World" - 2. Largo	9	4	**179**	1	0
Mahler: Symphony No. 5	8	11	**144**	0	0

Table 5.6: The results of the Sum of Absolute Difference distance-based classification process performed on 4-second split audio files, using rescaled data.

Composition	Breathiness	Brightness	Dullness	Roughness	Warmth
Bach: Brandenburg Concerto #3 - Allegro	1	1	**67**	0	0
Orff: Carmina Burana - O Fortruna	0	1	**29**	3	0
Rossini: The Barber of Seville: Overture	1	12	**71**	0	0
Smetana: Má Vlast - Vltava	2	10	**126**	0	0
Bach: Orchestral Suite #3 - Air	0	0	**51**	0	0
Fauré: Pavane	1	1	**60**	0	0
Gounod: Ave Maria	0	0	**31**	0	0
Mozart: Piano Concerto #21 - 2. Andante	0	1	**87**	0	0
Bach: Concerto For 2 Violins - Vivace	0	1	**43**	0	0
Brahms: Hungarian Dance #5	0	1	**31**	0	0
Grieg: In the Hall of the Mountain King	2	5	**26**	0	0
Grieg: Peer Gynt - Morning Mood	1	0	**44**	0	0
Beethoven: Moonlight Sonata: Adagio Sostenuto	6	1	**61**	0	0
Beethoven: Symphony #5 - Allegro Con Brio	0	0	**90**	0	0
Massenet: Thais - Meditation	0	9	**52**	0	0
Offenbach: The Tales Of Hoffman - Barcarolle	2	2	**39**	0	0
Chopin: Nocturne #2	0	3	**56**	0	0
Elgar: Pomp & Circumstance March #1	0	30	**39**	0	0
Mozart: Eine Kleine Nachtmusik: Allegro	1	1	**69**	0	0
Mozart: The Magic Flute - Overture	0	1	**76**	0	0
Giazotto: Adagio in G Minor	10	3	**90**	0	0
Mozart: Piano Sonata #11 - 3. Rondo Alla Turca	0	0	**41**	0	0
Mozart: Requiem Mass - Lacrimosa	1	0	**33**	0	0
Pachelbel: Canon	0	0	**42**	0	0
Barber: Adagio For Strings	4	19	**78**	0	0
Bizet: L'Arlésienne Suite #1 - Prelude	1	2	**80**	0	0
Sibelius: Finlandia	0	7	**90**	0	0
Strauss Jr. (J): The Beautiful Blue Danube	1	10	**60**	0	0
Beethoven: Fur Elise	2	1	**37**	0	0
Boccherini: String Quartet: Minuet	0	1	**45**	0	0
Corelli: Christmas Concerto - Allegro	0	3	**27**	0	0
Vivaldi: Violin Concerto In E, Op. 8:1, RV 269, "The Four Seasons (Spring)" - 1. Allegro	1	9	**29**	0	0
Bizet: Carmen Suite - Habanera	0	6	**21**	0	0
Mozart: Symphony No. 40	0	1	**84**	0	0
Strauss Sr. (J): Radetzky March	0	17	**20**	0	0
Wagner: Ride Of The Valkyries	0	21	**42**	0	0
Debussy: Suite Bergamasque - Clair De Lune	1	1	**61**	0	0
Myers: Cavantina	3	0	**45**	0	0
Rachmaninov: Vocalise	0	5	**65**	0	0
Tchaikovsky: Swan Lake Suite - Scene	0	1	**36**	0	0
Beethoven: Egmont Overture	1	3	**92**	0	0
Dvořák: Slavoni Dance No. 2	0	0	**67**	0	0
Satie: Gymnopedie #1	3	0	**41**	0	0
Verdi: Nabucco - Chorus Of The Hebrew Slaves	1	1	**58**	0	0
Beethoven: Symphony #9 In D, "Choral" - Finale: Ode To Joy	0	12	**19**	0	0
Handel: Messiah - Hallelujah Chorus	0	0	**43**	0	0
Holst: The Planets - Jupiter	0	7	**86**	0	0
Verdi: Medda de Requiem	0	3	**33**	11	0
Dvořák: Symphony #9, "From The New World" - 2. Largo	7	2	**145**	0	0
Mahler: Symphony No. 5	9	9	**112**	0	0

Table 5.7: The results of the Sum of Absolute Difference distance-based classification process performed on 4-second split audio files, using unscaled data.

implemented: SSD and SAD. Results of their testing, using similar methods as for the Euclidean distance have suggested similar conclusions to those of the Euclidean distance method. From these observations, it is evident that distance metrics may not be adapted to timbral values data, and therefore could not be used as a classification method.

Chapter 3 has introduced the field of AI research, which aims to achieve several goals imitating some human intelligence processes. In Section 3.3.2.2, different methods put forward by AI research have been discussed. One of these methods, named machine learning, is dedicated to solving classification issues by learning and creating classification models directly from data. Therefore, due to the different types of data retrieved from timbre characteristics, machine learning algorithms could be a solution to overcome this issue, which has been emphasised by the distance metrics results. The next section details the different machine learning approaches that have been applied to timbral values data estimated from audio files, in order to determine a timbre classification method.

5.5 Machine Learning Algorithms

This section presents the implementation of different machine learning algorithms developed to create classification models to automatically assign audio files into one of the five verbal attributes categories, based on calculated timbral values. Here, methods from the three main types of machine learning approach (i.e., unsupervised, supervised, and reinforcement learning) have been investigated in order to determine the best technique to handle timbral values data. The text starts with discussing the rationale for choosing machine learning algorithms for timbre classification. The development of each method is then detailed, along with reviews of their testing and performances, which served to determine the best approach for addressing the timbre classification challenge.

5.5.1 Motivations

Section 5.4 has presented the initial method for automatically classifying audio files of instrument combinations into one of the selected five verbal attributes, using estimations of specific timbre characteristics. Here, distance metrics have been utilised to determine the dominant perceptual quality by measuring the closeness of each calculated timbral value of an audio file with the corresponding known maximum timbral value. By comparing these values, the objective was to return the shortest distance as the dominant perceptual quality, due to it being the closest to an ideal value represented by the maximum value identified. However, from the observations of the results of different methods that have been implemented and presented in Section 5.4.3, it is evident that distance metrics do not harness the characteristics of timbral values data and thus, this method would not be appropriate for this specific classification purpose. Therefore, it was necessary to adopt a different classification approach.

Chapter 3 has introduced the field of AI research, presenting some of its objectives and approaches for addressing various challenges by harnessing human cognitive functions coupled to computers' processing abilities. In the discussions about the methods put forward by AI research (Section 3.3.2.2) it was suggested that machine learning algorithms were methods that can learn specific characteristics from data. Such methods are also often used for classification purposes, where classifier models are created from sets of training data. One of the benefits of using machine learning is that classification specifics do not need to be explicitly programmed, and therefore, it can be applied to complex types of data, which is the issue encountered with handling the data from timbre characteristics. Machine learning methods have been employed

in a broad range of applications and have proven to be effective techniques for classifying diverse types of data, as it has been discussed in Chapter 3.

The review of the use of AI methods in the musical domain, presented in Section 3.4, has highlighted that machine learning has also been applied in some musical applications for achieving various tasks and using different musical information. Furthermore, machine learning has also been used in research aiming to address some of the challenges faced in computer-aided orchestration, as mentioned in Section 2.5. For example, the *Live Orchestral Piano* system [172] utilised a machine learning approach for projective orchestration. Here, their machine learning algorithm has been trained to learn the properties of the probability distribution from a repertoire of different already made projective orchestrations, including works by Liszt who reduced Beethoven symphonies to piano for example. Then, the learning model allows their *Live Orchestral Piano* system to statistically predict, from an input piano score, projective orchestrations based on the learning material.

Following these observations, it arose that machine learning methods could be an appropriate approach for the present classification task using the data from specific timbre properties estimations. Therefore, different machine learning approaches have been implemented in order to determine the learning methods that could identify a classifier model from the timbral values. The next sections present the development of the different approaches, along with their testing and performances.

5.5.2 Unsupervised Learning

Machine learning regroups different approaches for identifying classification models from data. One of these methods is unsupervised learning, where machine learning algorithms aim to create a classifier model from 'raw' data, which means that datasets are unlabelled and the grouping of examples is unknown. This section details the motivations for investigating an unsupervised learning approach to create a classification model for timbral values from analysing audio files of instrument combinations. The text describes the dataset that has been created and utilised for the machine learning algorithm. Here, a k-means algorithm [252] has been selected, which will be detailed later in the section, along with results of its testing and performances.

5.5.2.1 Motivations

The advantage of using unsupervised learning methods is that an explicit function for classification, or categorisation of the data, is not required. Since there is no agreed metrics for timbre classification and the initial classification method presented in Section 5.4 did not perform successfully, unsupervised learning methods would, therefore, be a suitable solution to generate a function capable of grouping audio samples by analysing their calculated timbral values.

Machine learning algorithms require a training dataset to learn classification models, with sizes and types of datasets having an impact on the performances of the algorithms. In regards to unsupervised learning methods, training datasets need to be significant for the algorithms to be able to successfully identify classifier models, especially in the presence of complex data. It is evident that the learning algorithms accuracy increases proportionally with the quantity of examples. Therefore, the initial step is to establish a training dataset, which will then be processed by the machine learning algorithms. The next section details the development of the training dataset that has been input into an unsupervised learning method in order to create a classification function.

5.5.2.2 Dataset

As mentioned in the previous section, training datasets are an essential part of machine learning algorithms. In the case of unsupervised learning, large datasets are necessary due to the algorithm performing exploratory data analysis to identify classification models. Therefore, the first task is to create a training dataset, which will be used as input data for the machine learning algorithm. Here, timbral values of orchestrations generated by the *Orchids* program, used in the developments presented in Chapter 4, have been compiled and associated to the sets of 50 audio recordings of orchestral pieces utilised in Section 5.4.3. In order to have a larger training dataset, an analysis of an extra 205 audio recordings of varied orchestral pieces has been performed, using the same splitting methods as mentioned in Section 5.4.3 (i.e., audio recordings split into 1, 2, 3, 4, and 5 seconds audio files). This data gathering resulted in performing timbre estimations on 236,632 audio files, thus, compiling a training dataset composed of 236,632 values for each verbal attribute. Table 5.8 presents the statistics of the training dataset, which also highlights the disparity of the data for each timbral attribute.

Verbal Attribute	Number of samples	Minimum	Maximum	Range	Percentile	Median	Mean	Standard deviation	Variance
Breathiness	236632	0.0012	0.988	0.9868	0.4167	0.1836	0.2171	0.1467	0.0212
Brightness	236632	0.0237	0.9961	0.9724	0.6157	0.4155	0.421	0.1513	0.0229
Dullness	236632	0.0039	0.9773	0.9734	0.7706	0.5845	0.5791	0.1512	0.0229
Roughness	236632	0.0001	3017.2858	3017.2857	208.4811	17.339	72.0687	145.3299	21120.77
Warmth	236632	0.0001	0.1428	0.1427	0.025	0.0061	0.0102	0.0113	0.0001

Table 5.8: Statistics of the unscaled training dataset created from the analysis of 236,632 audio files.

5.5.2.3 *k*-Means Algorithm

A common unsupervised learning method is cluster analysis, where algorithms perform exploratory data analysis to identify hidden patterns or grouping in sets of data. Due to the unknown classification methods of timbral values, this approach offers the possibility of automatically learn a grouping from the data. Clustering models usually utilise similarity's measures based on Euclidean distance or probabilistic distance, which follows the approach described in Section 5.4. However, *k*-means algorithms utilise the mean of the data samples as its distance reference, while in the initial classification approach the distances were calculated from the maximum values. Here, using the mean, also named the centroid of the data, allows for identifying similarity from the centre of the data, instead of the extremes. The objective of the *k*-means algorithm is to divide a set of N samples X into K disjoint clusters, with μ_j as the centroid of the samples in the cluster. The *k*-means algorithm aims to select centroids that minimise the within-cluster sum of squared criterion, also called *inertia*, and is calculated as follows:

$$\sum_{i=0}^{n} \min_{\mu_j \in K} \left(\left\| x_j - \mu_i \right\|^2 \right) \tag{5.5}$$

where x_j is a data point and μ_i is the cluster centre (centroid).

The *k*-means algorithm has been implemented in `Python 3.5` using the `cluster.KMeans` function, taken from the `Scikit-Learn [253] v0.18` package[1], with the `n_clusters` parameter defined as 5, which informs the `cluster.KMeans` function to identify 5 clusters (representing the five verbal attributes) from the dataset. The function is also initialised with the `k-means++` methods, which speeds up convergence, meaning that the centroids μ_i are

[1] `http://scikit-learn.org/0.18/`

initialised distant from each other instead of applying a random centroid initialisation, providing better results [254]. Each entry of the dataset is input as a singular vector into the `cluster.KMeans` function, as follows:

$$x_i = [tBreathiness_i, tBrightness_i, tDullness_i, tRoughness_i, tWarmth_i] \tag{5.6}$$

where $tBreathiness_i$ is an audio file's calculated timbral value for the attribute *Breathiness*, $tBrightness_i$ for the attribute *Brightness*, $tDullness_i$ for the attribute *Dullness*, $tRoughness_i$ for the attribute *Roughness*, and $tWarmth_i$ for the attribute *Warmth*. The *k*-means clustering methods have been applied on the dataset using different rescaling methods, which are detailed in the following section.

5.5.2.4 Testing and Performance

This section presents the results and performances of the *k*-means algorithm implemented to identify a method to separate the set of samples into 5 clusters. The first step is to rescale the dataset in order to have values in the same ranges. Different rescaling methods have been applied to determine if it could have an impact on the performances of the clustering method.

The first rescaling method adopts the technique used in the initial classification approach, and described in Section 5.4.2. Here, timbral values are rescaled in the range 0.0–1.0 using the minimum and maximum values of each verbal attribute. The `cluster.KMeans` function has been applied to the 236,632 rescaled samples, with instructions to group the data samples into 5 categories. Due to the input data represented as a vector of 5 values, a Principal Component Analysis (PCA) algorithm has been applied on the clustering results. This method uses a Singular Value Decomposition (SVD) of the data in order to reduce its dimensional space. In this case, it enables the representation of the clustering results into a 2D space. Here, the Scikit-Learn's `decomposition.PCA` function has been utilised, and Figure 5.2 displays the *k*-means clustering results of the rescaled dataset.

The second rescaling method uses the Scikit-Learn's `MinMaxScaler` function, which scales samples of each verbal attribute in a given range (here 0.0–1.0). The rescaling method of this function is performed as follows:

$$x_{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \times (max - min) + min \tag{5.7}$$

where x_{min} is the minimum values in the set of samples, x_{max} the maximum values in the set of samples, and *min* and *max* the desired range's values. Here, the `cluster.KMeans` function has

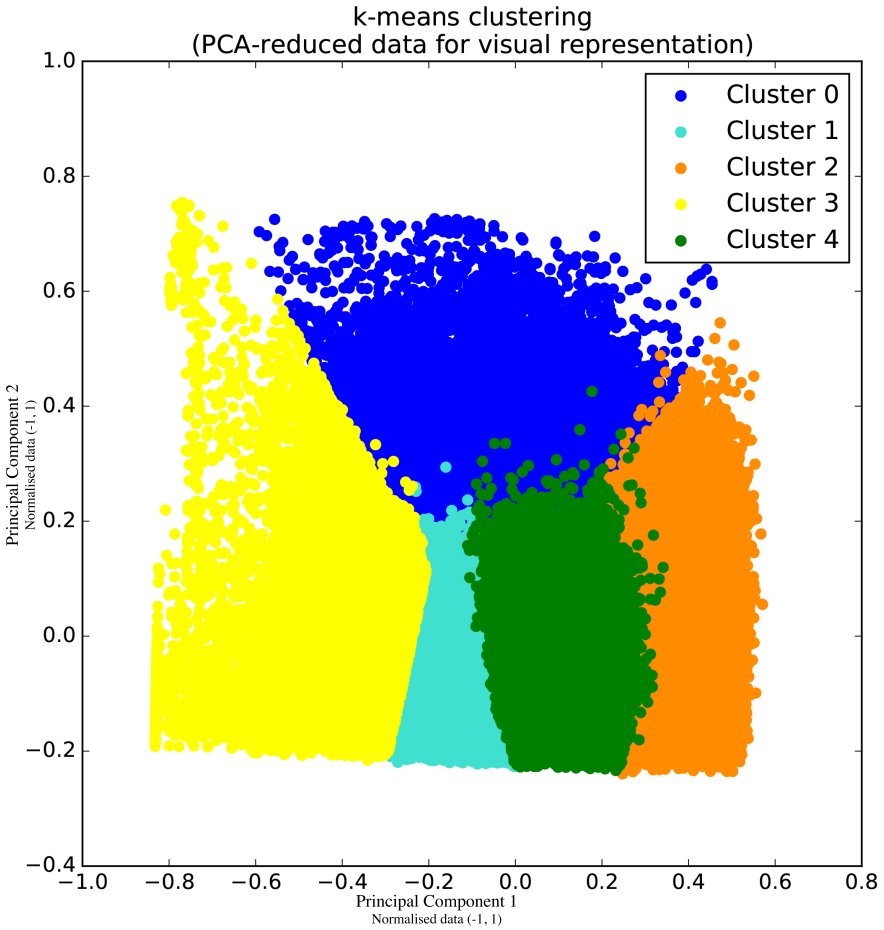

Figure 5.2: Graph showing the results for the *k*-means clustering performed on the 236 632 rescaled samples dataset.

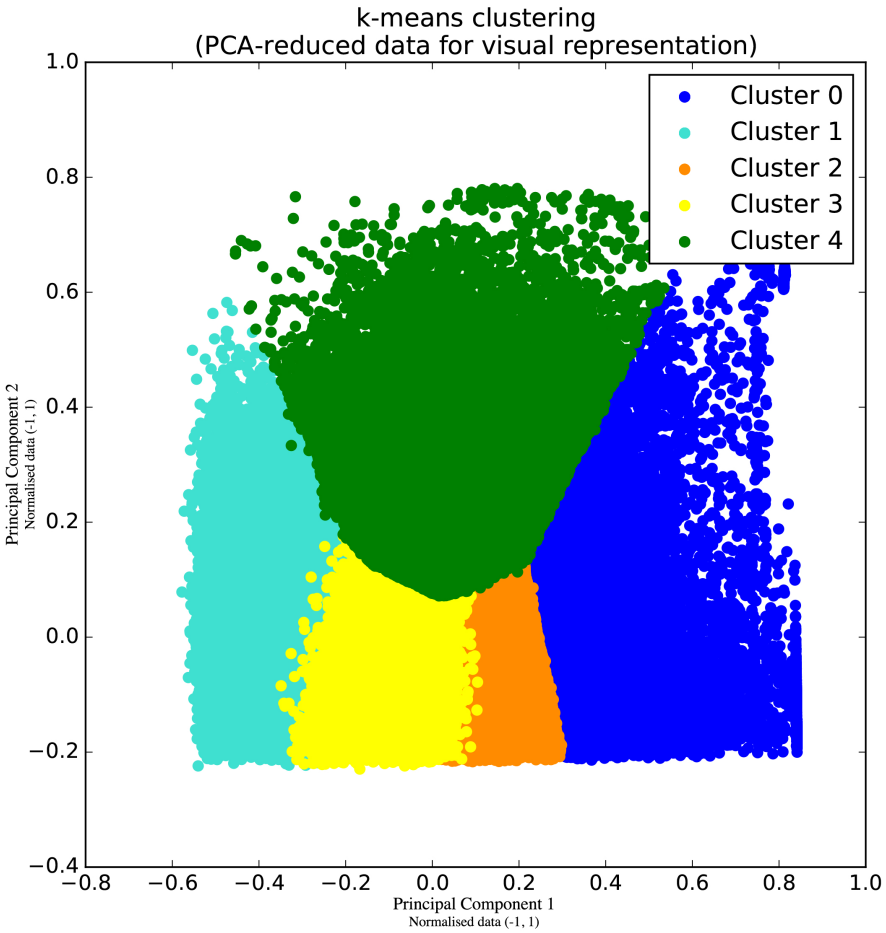

Figure 5.3: Graph showing the results for the *k*-means clustering performed on the 236,632 samples dataset, rescaled using the `MinMaxScaler` function.

also been applied to the 236,632 rescaled samples in order to create 5 clusters from the data. Again, the Scikit-Learn's `decomposition.PCA` function has been used to represent the results of the *k*-means clustering performed on the rescaled training dataset. Figure 5.3 displays the results for the dataset rescaled using the Scikit-Learn's `MinMaxScaler` function.

Finally, the third rescaling method is performed by the Scikit-Learn's `MaxAbsScaler` function. Here, the timbral values for each verbal attribute are scaled using the corresponding maximum absolute value, and the dataset is rescaled in the range -1.0–1.0. This method does not center the data, which aims to keep any sparsity in the data. Again, the Scikit-Learn's `decomposition.PCA` function has been used to represent the results of the *k*-means clustering performed on this rescaled training dataset into a 2D space. Figure 5.3 displays the results for

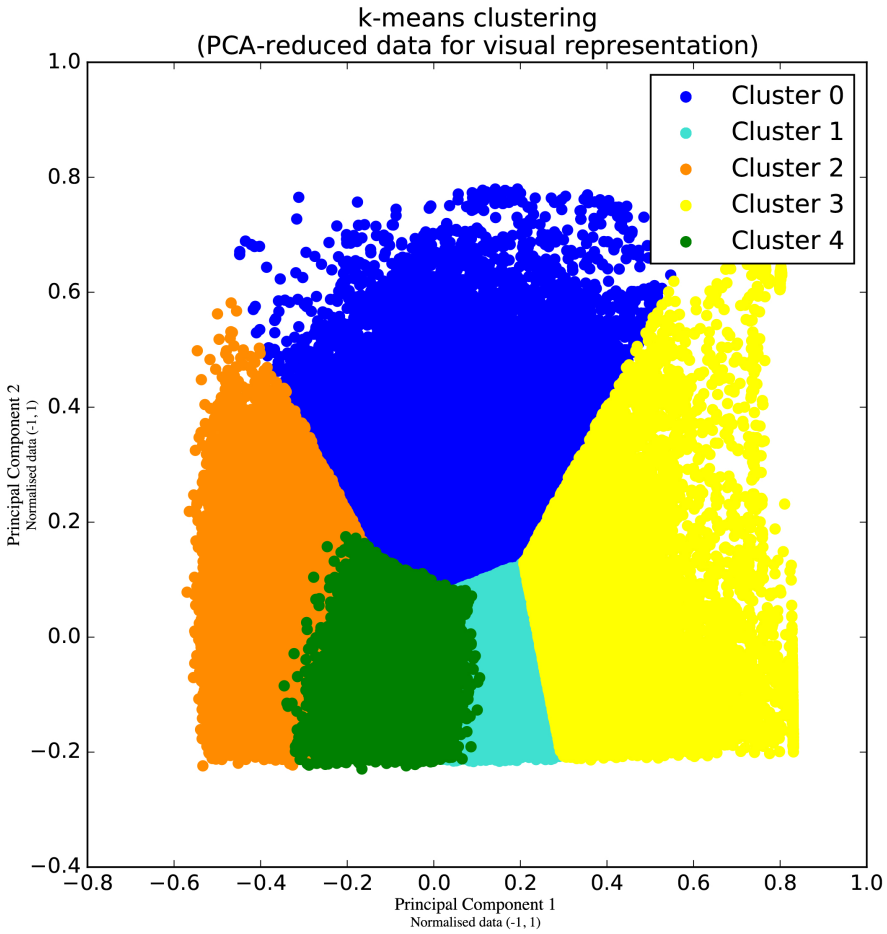

Figure 5.4: Graph showing the results for the *k*-means clustering performed on the 236,632 samples dataset, rescaled using the `MaxAbsScaler` function.

the dataset rescaled using the Scikit-Learn's `MaxAbsScaler` function.

5.5.2.5 Discussions

The previous section has detailed the different techniques that have been employed to rescale the training dataset, on which a *k*-means algorithm has been applied. Here, the clustering method was designed to divide the dataset into 5 categories, representing the five verbal attributes. The results of the clustering applied on each of the three rescaled training datasets, displayed respectively in Figures 5.2, 5.3, and 5.4, indicate similar shapes for the 5 clusters identified from the data. However, due to the data being unlabelled, these results do not provide any information about which cluster corresponds to which verbal attribute.

To test the output of the classification model proposed by the *k*-means algorithm, a set of testing samples have been created by manually labelling randomly selected samples taken from the training dataset. Figure 5.5 shows the normalised confusion matrix for the testing of the *k*-means classification model created from the rescaled training dataset. Here, each cluster has been assigned randomly to a verbal attribute, because of the unknown repartition of the clusters. 26 manually labelled testing samples for each verbal attribute have been input into the classifier model, which then predicted the category of each sample from analysing its timbral values. The confusion matrix plots the predicted labels, on the horizontal axis, against the true labels of the testing samples on the vertical axis. Results of the confusion matrix suggest that the classification model proposed by the *k*-means clustering method is able to predict the category of a sample by only analysing its five timbral values. This is supported by the high values of the confusion matrix scores. However, these results also indicate that the clusters were not assigned the correct verbal attributes. For example, in Figure 5.5, the testing samples labelled *roughness* were predicted as samples from the category *breathiness* by the classification model. Figure 5.6 shows the normalised confusion matrix for the *k*-means classification model created from the training dataset rescaled with the Scikit-Learn's `MinMaxScaler` function, and Figure 5.7 from the training dataset rescaled with the `MaxAbsScaler` function. The results represented in these two confusion matrices suggest similar classification successes, with also some mislabelling for the clusters.

The results of the testing of the *k*-means clustering method support the approach of grouping data into 5 categories, which represent the five verbal attributes. They also suggest that the different rescaling techniques do not have a significant impact on the classification models as indicated by the values of the confusion matrices displayed in Figures 5.5, 5.6, and 5.7. Nevertheless, the *k*-means clustering method involves an extra step, which consists of evaluating the suggested clusters in order to assign the verbal attributes categories to their correspondent clusters. This task requires the listening of audio files and manually evaluating and labelling the samples, which are then input into the classification model. Following the requirement of a labelling samples task, it was decided to investigate methods from the supervised learning category, which operate with labelled training dataset, in order to evaluate the classification models suggested by such methods. The next section presents the supervised learning algorithms that have been implemented, along with the results of their testing and performances.

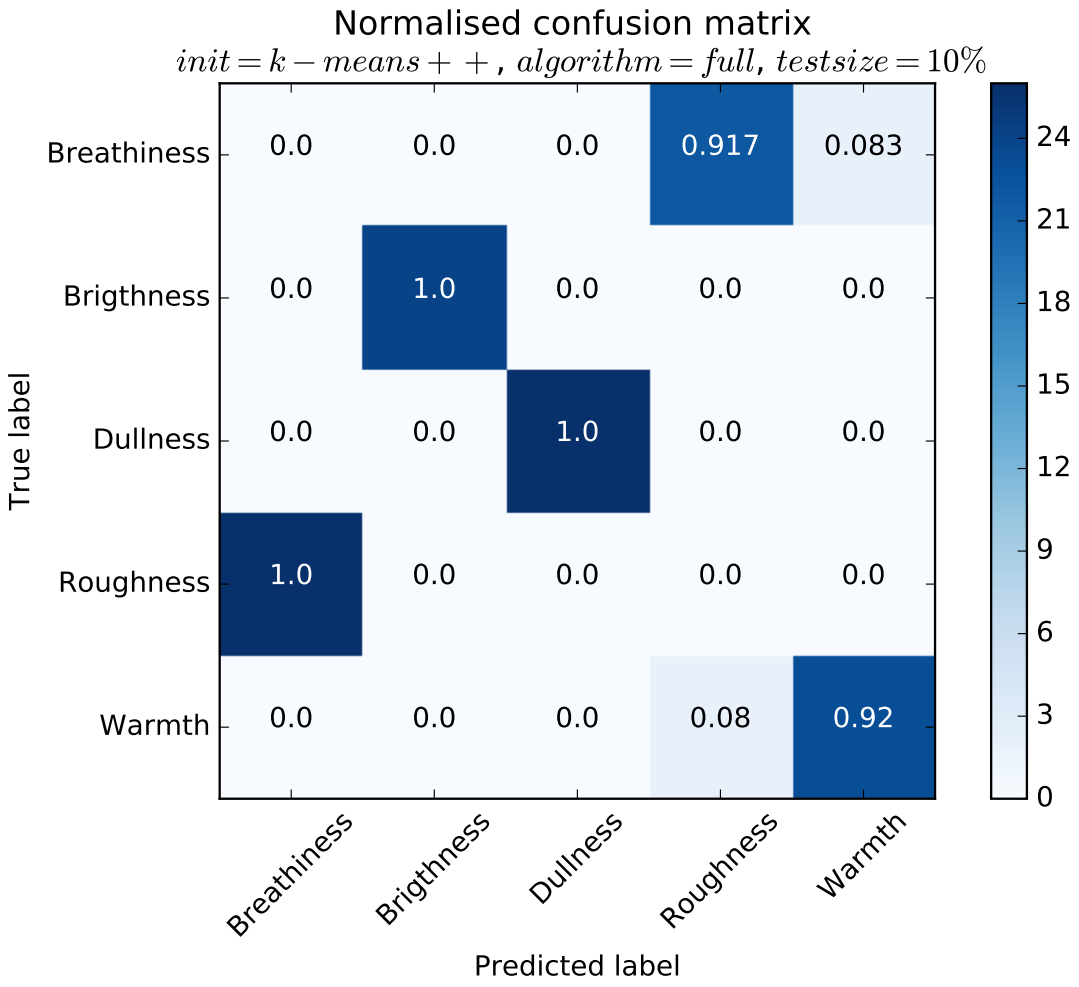

Figure 5.5: Normalised confusion matrix for the testing of the *k*-means clustering model identified from the 236,632 rescaled samples dataset. 26 testing samples for each of the verbal attributes have been used, and clusters assigned randomly to a verbal attribute.

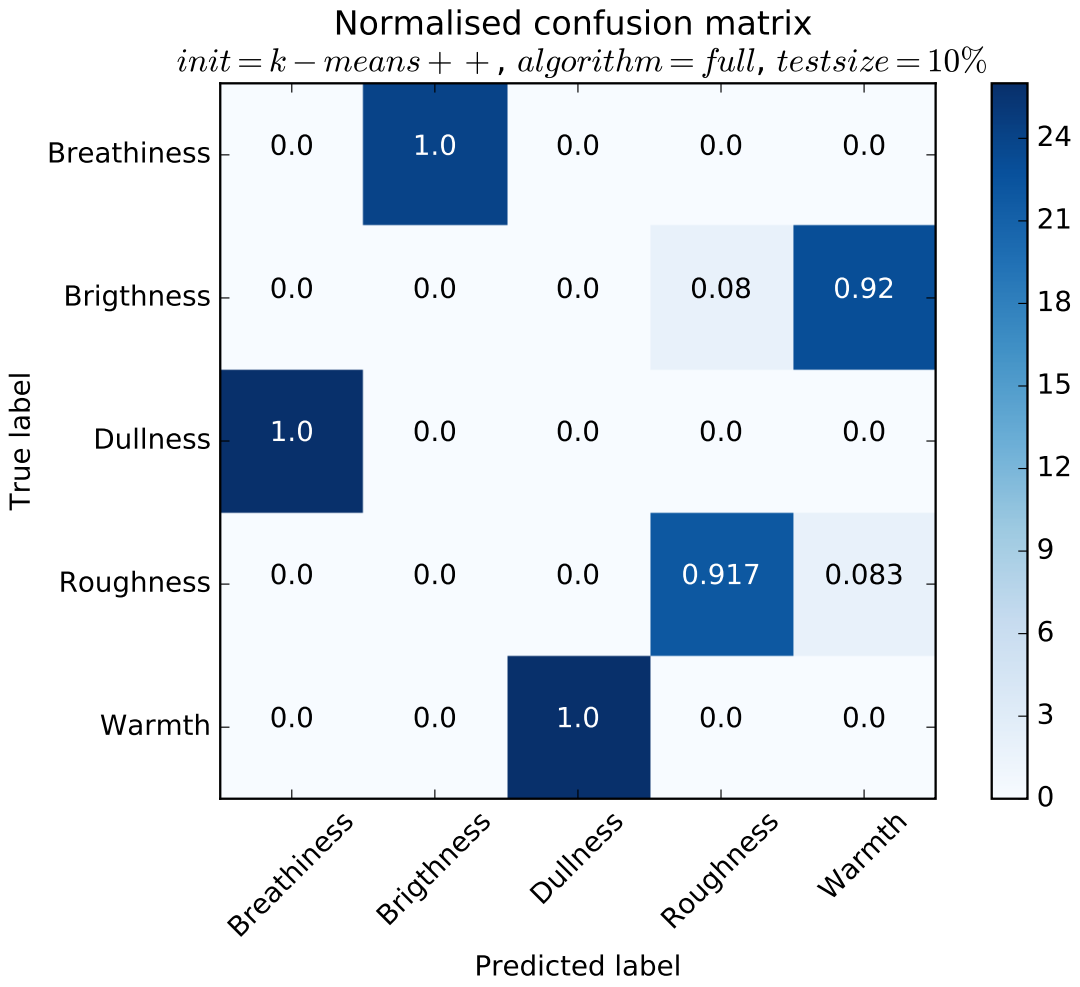

Figure 5.6: Normalised confusion matrix for the testing of the k-means clustering model identified from the 236,632 samples dataset, rescaled using the `MinMaxScaler` function. 26 testing samples for each of the verbal attributes have been used, and clusters assigned randomly to a verbal attribute.

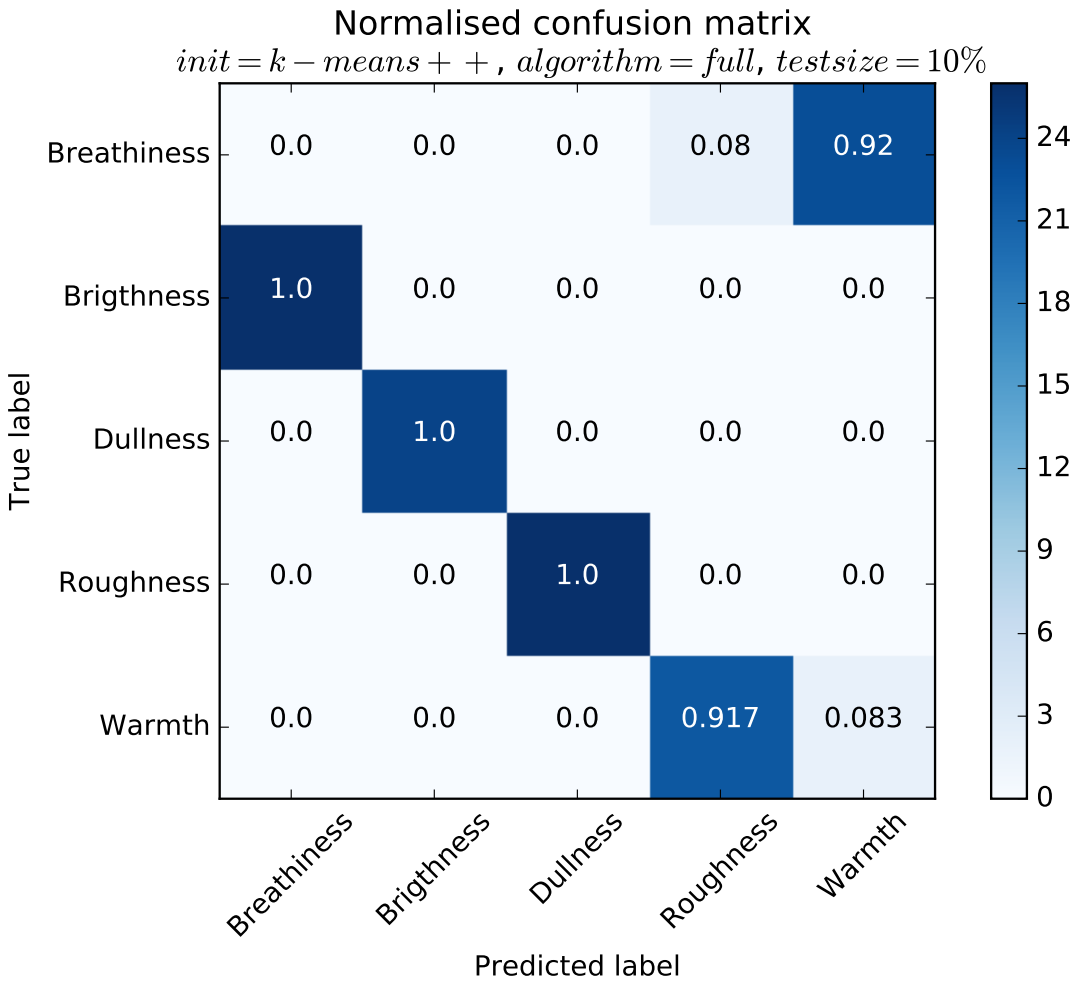

Figure 5.7: Normalised confusion matrix for the testing of the *k*-means clustering model identified from the 236,632 samples dataset, rescaled using the `MaxAbsScaler` function. 26 testing samples for each of the verbal attributes have been used, and clusters assigned randomly to a verbal attribute.

5.5.3 Supervised Learning

This section presents the investigations of using supervised learning methods for the classification of audio files based on their timbral values. Machine learning algorithms from this category aim to generate classifier models from datasets containing training examples. Here, training datasets need to be labelled, which means that each training sample consists of the input values and the label of the expected category in which it should be classified. The text details the motivations for investigating supervised learning methods and the creation of a training dataset. Then, the implementation of two supervised learning algorithms is described, along with the results of their classification performances.

5.5.3.1 Motivations

The results of the classification models created by the *k*-means clustering method (Section 5.5.2) have suggested that an unsupervised learning algorithm could be used to learn classifier functions from sets of timbral values. The *k*-means algorithm has been able to create five clusters, which represent the five verbal attributes. However, due to the unlabelled training data, it was not possible to identify the classification categories without evaluating the clusters by inputting manually labelled samples into the classification models. This extra task involves creating a labelled testing dataset, which could also be used as labelled training data for supervised learning techniques. Therefore, it would be interesting to investigate also the performances of machine learning algorithms designed to create classification models based on a set of examples, instead of learning the classification from large sets of data.

Two different supervised learning algorithms have been evaluated: Support Vector Machine (SVM) and Artificial Neural Networks (ANNs). These two methods can be trained with examples from which they are able to identify a classification model and then predict the categories of new inputs. Specifics of these two methods are detailed in Sections 5.5.3.3 and 5.5.3.6 respectively. However, before implementing and evaluating these two algorithms, it is necessary to create a labelled training dataset, which is detailed in the next section.

5.5.3.2 Training Corpus

As mentioned above, supervised learning algorithms are dependent on a labelled training dataset. Therefore, the initial step is to create a set of examples that will then be used to train the machine learning algorithms. In this case, examples consist of audio files' calculated

timbral values as input data and their dominant perceptual quality, represented by the verbal attributes, as the desired output values. The training samples have been selected from the large training dataset created for the unsupervised learning method, described in Section 5.5.2.2. Here, the 250 audio files that have scored the highest values for each verbal attribute have been selected and labelled with their corresponding attribute. For example, the 250 samples with highest values for the attribute *brightness* have been chosen and labelled 'brightness'. In total, the corpus training contains 1250 labelled samples, which can be utilised for evaluating the two supervised learning algorithms detailed in the following sections.

5.5.3.3 Algorithm 1 - Support Vector Machines (SVM)

The first supervised learning algorithm that has been implemented is Support Vector Machines (SVM) [202], a method often utilised for data classification, which is the required task here, and also for regression analysis. This machine learning method has been successfully applied in various applications, such as face detection [255] and speaker recognition [256] to name but two. SVM models try to find the separation, also called the hyperplane, between the different categories with a gap that is as wide as possible to create a delimited space for each category. Then, when a new value is presented, SVM models estimate the space in which the value sits, thus predicting the category it belongs to. Considering the randomly generated points shown in Figure 5.8a, where points fit in two categories (i.e. blue or red), the SVM algorithm will seek to divide the space into two distinct areas representing the two categories. One possible classification model is shown in Figure 5.8b, where the categories are delineated by a straight line. Here, the decision function, also called kernel, uses a linear model. Other decision functions can be used to optimise the separation, such as the classification models shown in Figures 5.8c and 5.8d, respectively, using a polynomial kernel and a radial basis function kernel. Here, we can note that using these kernels there is no classification error, while using the linear model shown in Figure 5.8b, one blue point was in the red space.

The SVM algorithm has been implemented in `Python 3.5` using the `svm.SVC` function, which is a SVM method for classification task, taken from the `Scikit-Learn v0.18` package. Here, for the initial implementation, the `svm.SVC` function's default parameters have been used, which are kernel type $= rbf$ (for Radial Basis Function), rbf kernel coefficient $\gamma = 1/number\ of\ samples$, and penalty parameter $C = 1.0$. The kernel type defines the decision

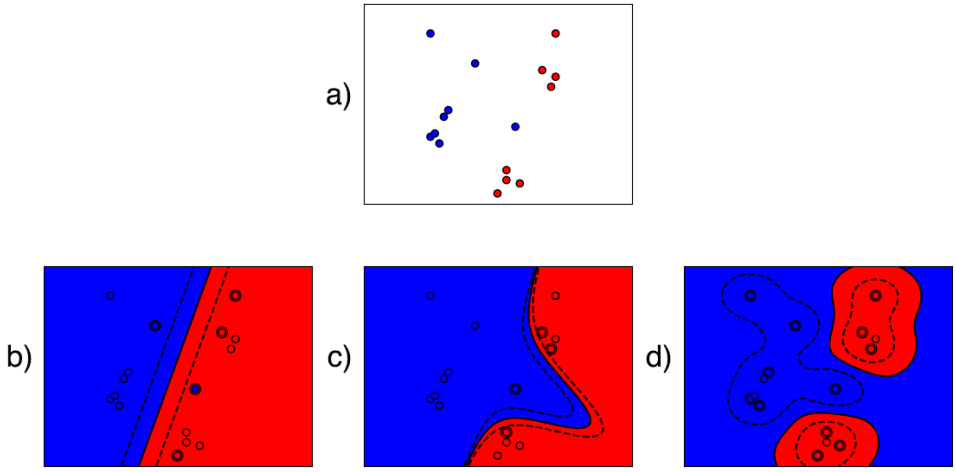

Figure 5.8: a) is a graph representing random points that belong into two categories (i.e. blue and red). b) shows a classification model for the data point shown in a), created by a SVM algorithm with a linear kernel. c) is classification model using a polynomial kernel, and d) used a radial basis function kernel.

function to be used for the delimitation of the categories. The radial basis function is non-linear, meaning that categories are not delimited by a straight line but in this case by radial axes (Figure 5.8d). This decision function for a RBF kernel is mathematically defined as:

$$\exp\left(-\gamma\left\|x-x'\right\|^2\right) \tag{5.8}$$

where γ is specified and greater than 0, and x is a set of points. The default parameter γ for the `svm.SVC` function is set as $1/number\ of\ samples$. The penalty parameter C determines the tolerable misclassification errors for the SVM function optimisation. In other words, the number of points that can be fitted outside of their category in the search for the delimitations. Further details about specific `Scikit-Learn`'s `svm.SVC` parameters can be found in [257].

The SVM model is then trained using the training corpus described in Section 5.5.3.2. Here, one training example consists of a vector for the timbral values as input data and the identified timbral descriptor as the category's label. Testing and performance of the classification model suggested by the SVM method is detailed in the next section.

5.5.3.4 Testing and Performance

To test the performances of the classification model learned by the SVM method, the training corpus has been divided into a set of training samples and a set of testing samples. Similarly to the testing of the *k*-means clustering algorithm, different rescaling techniques have been applied onto the timbral values dataset before inputting it into the SMV algorithm. Figure 5.9 shows the normalised confusion matrix of the SVM classification model using samples rescaled following the initial technique described in Section 5.4.2. Here, the training samples consisted of 90% of the training corpus (1125 samples), and 10% of the training corpus (125 samples) were selected as testing samples. Using this training dataset, the `svm.SVC` function scored a success rate of 0.976, which means that the classification model predicted the correct verbal attribute 97.6% of the time. Similar testing methods, using the default `svm.SVC` function, have been applied to the corpus training rescaled using `Scikit-Learn`'s `MinMaxScaler` function, and `Scikit-Learn`'s `MaxAbsScaler` function. The classification model also scored a success rate of 0.976 and produced similar confusion matrices using the training corpus rescaled with both techniques, which suggests that the different rescaling methods do not impact the learning process.

The performances of the SVM classification model have also been tested using the unscaled training corpus. Figure 5.10 displays the normalised confusion matrix, also using the 90% training samples–10% testing samples division. Using this training dataset, the `svm.SVC` function scored a success rate of 0.872, which is lower than the score of the classification model created from the rescaled training dataset. This indicates that the supervised learning process performs better with the rescaled timbral values.

5.5.3.5 Parameter Tuning

The success rate of the classification model created by the `svm.SVC` function suggested that SVM could be utilised for the classification of audio files using their timbral values. The high success rate has been obtained using only the default parameters of the `svm.SVC` function. Therefore, in order to optimise the classification success rate, a parameter tuning process has been performed to obtain the best parameter combination. Here, the `Scikit-Learn`'s `model_selection.GridSearchCV` function has been utilised to test and compare the performances of different combinations of parameters. This process has determined the best SVM parameters as kernel type $= rbf$, penalty parameter $C = 10$, and rbf kernel coefficient $\gamma = 0.001$,

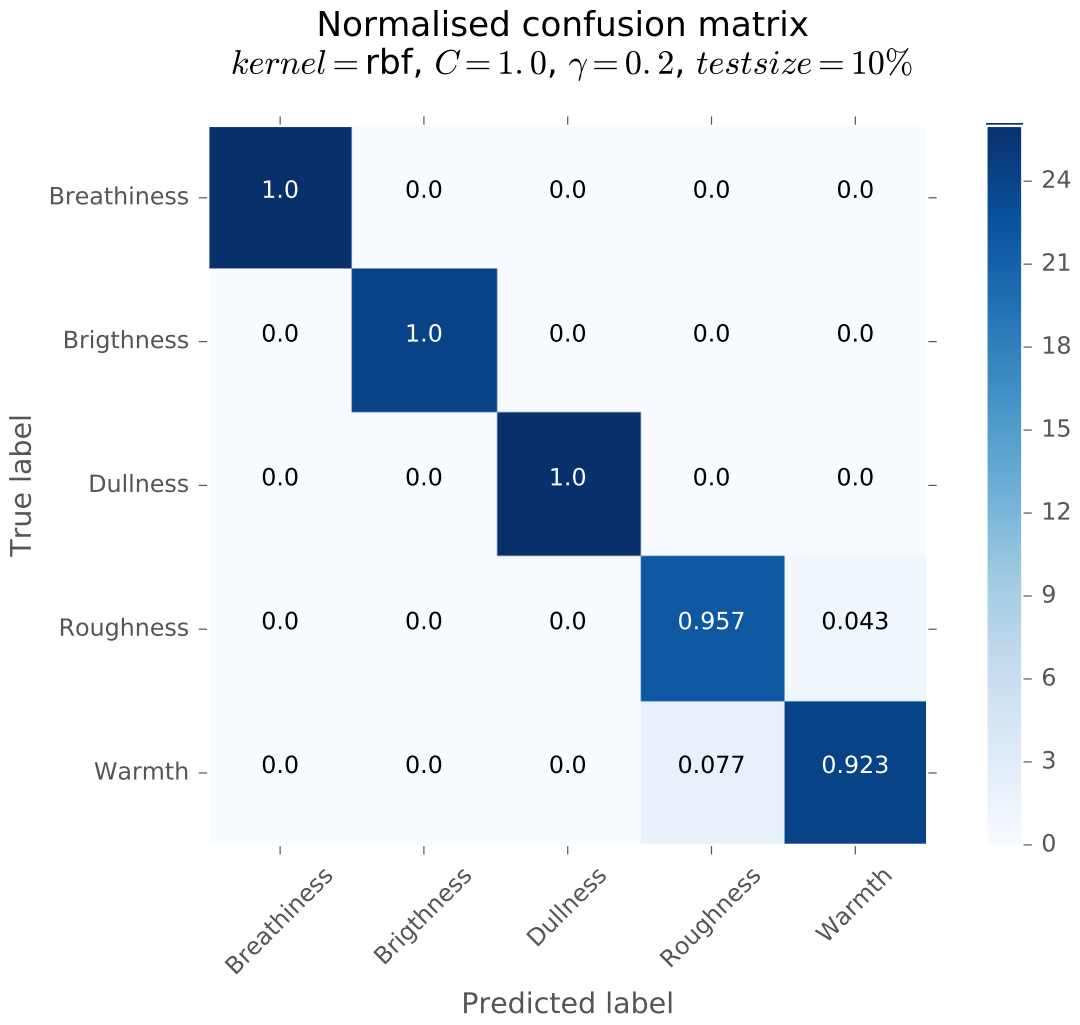

Figure 5.9: Normalised confusion matrix of the SVM classification model created from rescaled training dataset, with test size = 10% (1125 training samples, 125 testing samples).

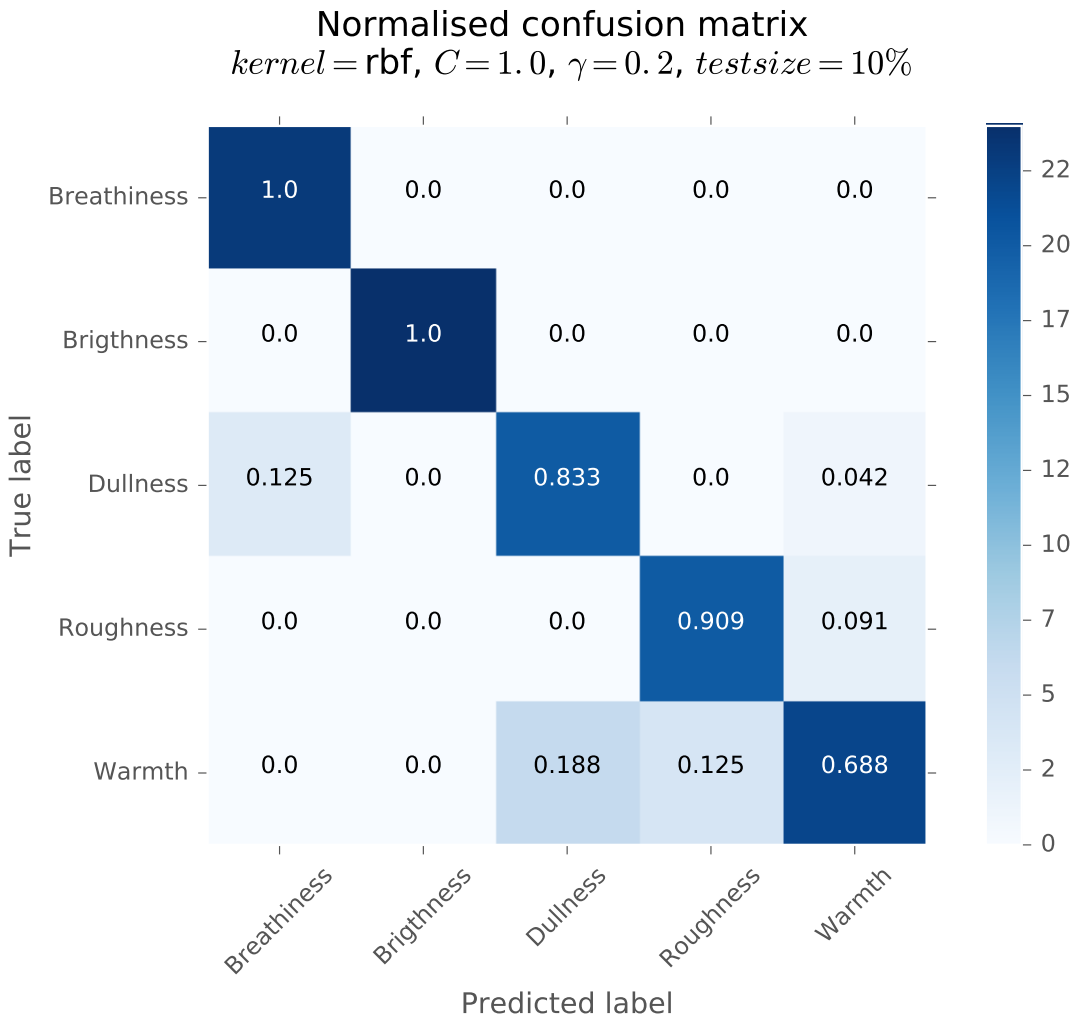

Figure 5.10: Normalised confusion matrix of the SVM classification model created from unscaled training dataset, with test size = 10% (1125 training samples, 125 testing samples).

which scored a success rate of 0.978, a slightly better score than with the default parameters. Nevertheless, this result indicates that an RBF kernel is the most appropriate type for this classification task. Different details about the parameter tuning of the `svm.SVC` function can be found in Appendix A.

5.5.3.6 Algorithm 2 - Artificial Neural Networks (ANNs)

The second supervised learning algorithm that has been implemented is Artificial Neural Networks (ANNs), which are non-linear statistical data modelling algorithms that have been used in many applications, such as speech recognition [258], object recognition [259] and also for musical tasks (Section 3.4).This method is inspired by the structure of biological neural networks, such as the human brain. Here, neurons are interconnected and information passing through can change their states based on the input and output [203, 204]. A model of an artificial neuron, called a perceptron, is shown in Figure 5.11. Here, we have a single layered neural network where the neuron sums all the input information, which can have weighting values, and its result is passed through the activation function, which will decide the output. For example, if the input's summation is greater than a defined threshold, the neuron will output 1. ANNs typically consists of many neurons with multiple layers and aim to learn a function $f : X \rightarrow Y$ using the backpropagation method by analysing training examples. In other words, we input data points x_i with their corresponding labels y_i and the ANNs algorithm will update its neurons from this information to learn the function f that associates x_i to y_i. In this investigation, the x_i were vectors of timbral values and y_i were their corresponding timbral descriptors.

The ANNs algorithm has been implemented in `Python 3.5` using the `neural_network.MLPClassifier` function, which is a multi-layer perceptron classifier method, taken from the `Scikit-Learn v0.18` package. Similar to the SVM implementation, the default values of the `neural_network.MLPClassifier` function have been used. These values are 100 neurons in hidden layers, $activation = relu$ (rectified linear unit function) and $solver = adam$. Here, the ANNs consisted of hidden layers composed of 100 neurons, with the activation function based on rectified linear unit (ReLU), defined as:

$$f(x) = max(0, x) \tag{5.9}$$

where x is the input to a neuron. The solver function *adam* (short for Adaptive Moment Estimation) is a stochastic gradient-based optimiser for the input weighting values w_i and is based

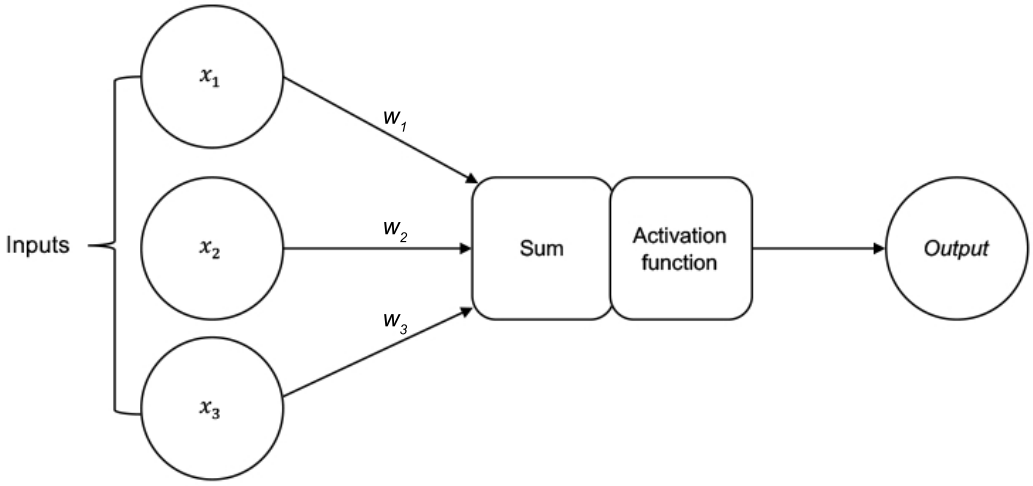

Figure 5.11: Model of an artificial neuron called a perceptron.

on [260].

The training corpus detailed in Section 5.5.3.2 have been utilised as input values to train the ANNs classifier algorithm. Similar to the SVM implementation, each training example consisted of a vector of timbral values and the corresponding timbral descriptor as the desired output. The next section details the testing method and performance of the classification model suggested by the ANNs method.

5.5.3.7 Testing and Performance

Similar testing methods used for the SVM algorithm, and described in Section 5.5.3.4, have been followed in order to test the classification model created by the ANNs algorithm. Here, the training corpus has also been divided into batches of training samples and testing samples. Figure 5.12 shows the normalised confusion matrix of the ANNs classification model using samples rescaled following the initial technique described in Section 5.4.2. In these results, the dataset consisted of 90% of the training corpus (1125 labelled samples), and 10% of the training corpus (125 samples) were selected as testing samples. Using this training dataset, the `neural_network.MLPClassifier` function scored a success rate of 0.976, which is similar to the score the SVM classification model presented previously. Again, testing using the dataset rescaled with Scikit-Learn's `MinMaxScaler` and `MaxAbsScaler` functions have been performed, which both scored the same success rate of 0.976. Classification model using un-

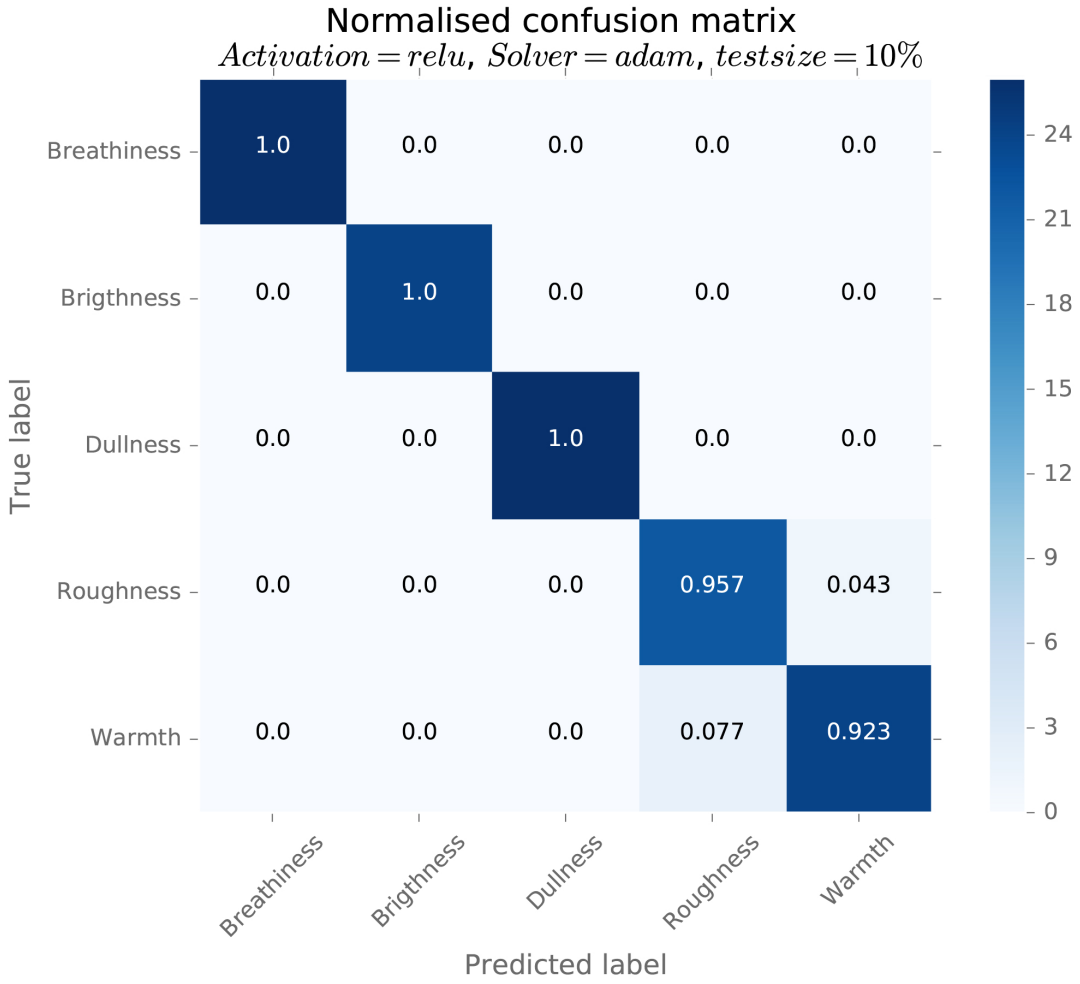

Figure 5.12: Normalised confusion matrix of the ANNs classification model created from rescaled training dataset, with test size = 10% (1125 training samples, 125 testing samples).

scaled training corpus has also been tested. Figure 5.13 shows the normalised confusion matrix of the ANNs classification model created from the unscaled training dataset, which produced a success rate of 0.88, significantly lower than with rescaled data. This indicates that the ANNs supervised learning process also performs better with the rescaled timbral values.

5.5.3.8 Parameter Tuning

Similarly to the SVM implementation, a parameter tuning process has been performed to obtain the best ANNs parameter combination. Again, the `Scikit-Learn`'s `model_selection.`

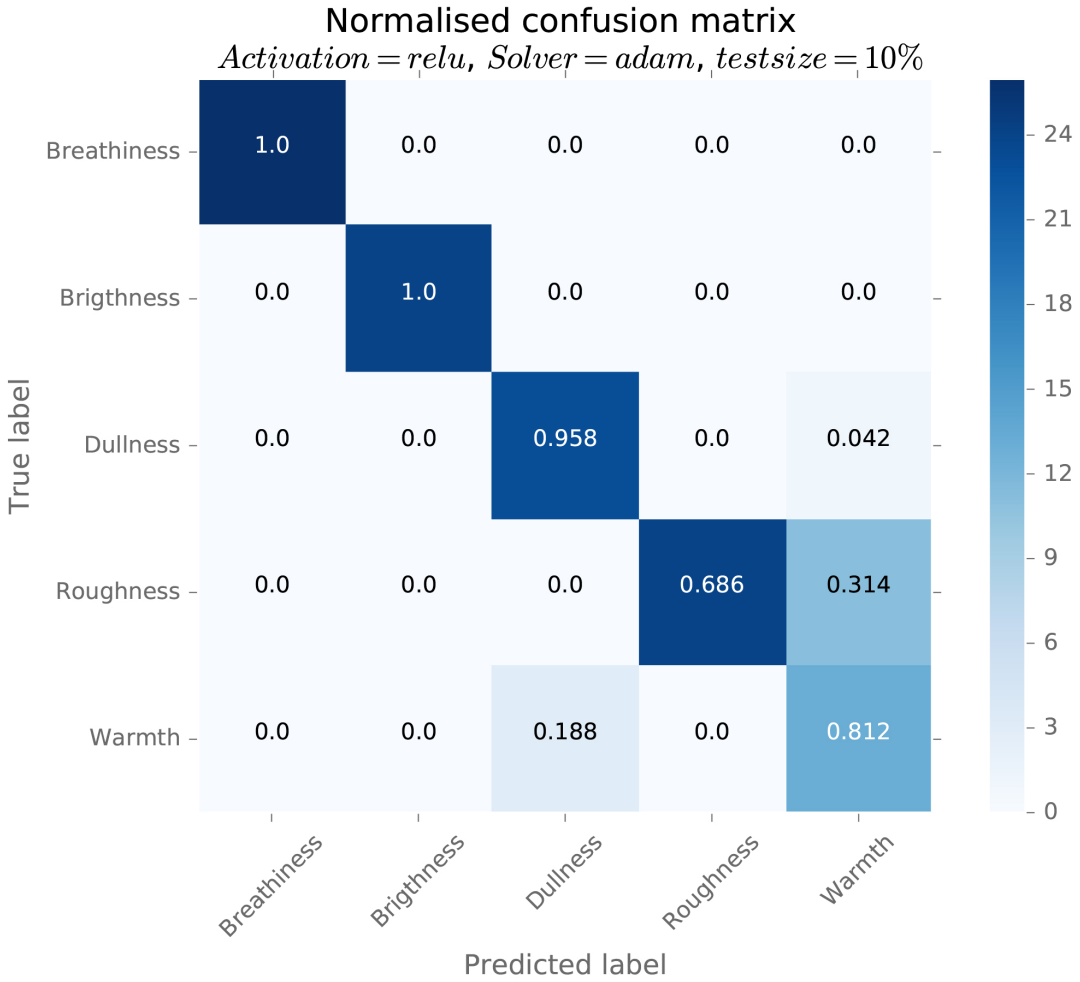

Figure 5.13: Normalised confusion matrix of the ANNs classification model created from unscaled training dataset, with test size = 10% (1125 training samples, 125 testing samples).

`GridSearchCV` function has been utilised to test and compare the performances of different combinations of `neural_network.MLPClassifier` function parameters. The best classification success rate has been obtained with the parameters $activation = identity$, and $solver = adam$. Here, the activation function is defined as $f(x) = x$. Using these parameters, the ANNs classification model scored a success rate of 0.984. Different details about the parameters tuning of the `neural_network.MLPClassifier` function can be found in Appendix A.

5.5.3.9 Discussions

This section has presented the implementation of two supervised learning algorithms. Classification models suggested by these two approaches have been tested by dividing the corpus training into batches of training samples and testing samples. A parameters tuning process has been applied to both functions to obtain the parameter combination producing the best classification score. Here, the SVM classification model scored a success rate of 0.978, with the ANNs classification model achieving a slightly better success rate of 0.984, using one batch of 1125 training samples and one consisting of 125 testing samples. These high success rates suggest that classification models generated by these two supervised learning algorithms are able to identify the perceptual qualities from audio files.

Although these two supervised learning algorithms present high success rates, they are heavily dependent on a corpus training labelled beforehand. This task involves the process of listening to audio samples and manually labelling their perceptual qualities. While this required action can be seen as a negative extra task, it could be used to allow the users to create their own training dataset, which would, therefore, represent their own perceptual preferences. This could also offer a solution to overcome the challenge of the variation in perception levels between individuals. The next section presents a method to include this user's personalisation into a supervised learning algorithm.

5.5.4 Reinforced Supervised Learning

The previous sections have detailed the successful classification models generated by the supervised learning algorithms. As indicated, such methods are dependant on the creation of a labelled training dataset, which is an additional task that some would prefer to avoid. However, in this case, the manual labelling task could be used to calibrate the classification model. The selected approach is inspired by the reinforcement learning methods [261], where classification

models are generated from evaluations of outputs. Within this approach, suggested classifications are evaluated using a system of reward–penalty, which information is then processed by the machine learning algorithm to calibrate the model. However, in the present investigation, the selected approach uses a supervised learning algorithm instead of a bespoke reinforcement learning algorithm. The following sections detail the rationale for selecting this reinforced supervised learning approach, along with descriptions of the training dataset and technical details about the implemented algorithm.

5.5.4.1 Motivations

The developments about the two supervised learning algorithms have suggested that this approach could be a viable method to identify the perceptual quality of an audio file by analysing its timbral values, which is supported by the high success rates of 0.978 for the SVM classification model and 0.984 for the ANNs. As mentioned previously, these methods rely on a training dataset that has been previously labelled, as classification models are learned from sets of examples. It is evident that this listening and labelling process represents an extra task in the creation of a classification model. However, this process could be used to address the challenging variation of perception between individuals. Here, by evaluating different audio files, users could calibrate the classification models to their own musical perception. The supervised learning methods are utilised to generate standard perception classification models, while a reinforced supervised learning method could offer user-specific perception classification models. The development of such an approach is detailed in the following sections.

5.5.4.2 Training Dataset

Similarly to the supervised learning methods, a training dataset needs to be created in order to generate a standard classification model. Following the high classification scores suggested by the supervised learning algorithms, the same training dataset, described in Section 5.5.3.2, has been selected. Here, each verbal attribute contains 250 training samples. For the reinforcing approach, a weighting method has been selected, which will be used for the reward–penalty mechanism. Thus, a weight value has been added to each training sample, with 50 considered as the standard value. Furthermore, a function is designed to select a certain number of audio files from the training dataset, with the number of files being defined by the user, and it presents the audio files to the user in order to be evaluated. Here, the users can listen to the audio

files, and then label the verbal attribute that corresponds to their own perception of the sound. These newly labelled samples are then added to the training dataset, with their correspond weight values set at 100. This listening and labelling process can be repeated multiple times if required, which further calibrates the classification model to a user-specific perception. The next section describes the technical details of the reinforced supervised learning method.

5.5.4.3 Algorithm - SVM with Weighted Samples

The reinforced supervised learning approach uses a SVM algorithm. Here, this method has also been implemented in `Python 3.5`, using the `Scikit-Learn v0.18`'s `svm.SVC` function, with penalty parameter $C = 10$, kernel type $= rbf$, and rbf kernel coefficient $\gamma = 0.01$ as suggested by the parameters tuning process presented in Section 5.5.3.5. Whereas in both supervised learning algorithm presented previously the input data consisted of the timbral values and the verbal attribute for each sample, here the weight value of each sample is input as well. These weight values are processed by the `svm.SVC` function in order to emphasise the training samples with highest values for the generation of the classification model.

5.5.4.4 Discussions

This section has presented the implementation of a reinforced supervised learning algorithm. This approach is built on the successful classification models generated by the supervised learning methods presented in Section 5.5.3. It is also inspired by the reinforcement learning methods, where models are calibrated by evaluating their outputs, using a reward–penalty technique. However, here, a weighting samples approach has been chosen, where the evaluations change the value of the samples' weight, thus calibrating the classification model.

The reinforced supervised learning algorithm is built on the SVM algorithm described in Section 5.5.3.3, which has scored a classification success rate of 0.978 using the methods detailed in Section 5.5.3.4. For the reinforcement part, a weight value has been added to each sample of the training dataset detailed in Section 5.5.3.2, which is then used to inform the classification model on which samples to emphasise. Here, the weight values can be altered directly in the labelled training dataset, or by performing a listening and evaluation process. The first technique offers the ability to create a personal weighted training dataset, while the second calibrates the standard training dataset to a user-specific perception model.

Following the successful classification models generated by the supervised learning methods, the approach presented in this section has proposed a solution to overcome the challenge

of the variation in perception levels between individuals by offering the ability to calibrate the classification models. This approach, inspired by the reinforcement learning methods, allows the users to input their own perception levels into the learning process, thus improving the accuracy of the suggested classifications.

5.6 Chapter Conclusions

The investigation presented in this chapter has been built on the findings suggested in Chapter 4, which has identified methods for calculating specific timbre properties, represented by verbal attributes, from sounds being created by combining diverse instruments. One of the limitations of the timbral ranking system presented in the previous chapter was that it outputs 'top' result, even if none of the audio files had the characteristics of a perceptual quality, which was due to the comparison being performed between the audio files present in a folder. In order to overcome this limitation, methods for classifying audio files according to specific perceptual qualities have been studied.

Similar verbal attributes and methods of calculation put forward by the timbral ranking system presented in Chapter 4 have been selected for this investigation, which are described in Section 5.3. Performing these estimations resulted in obtaining a timbral value for each attribute, which could then be utilised in a classification method. Thus, an initial classification approach has been studied, which was based on distance measures. Due to a lack of agreed metrics, a timbral analysis has been performed on several orchestral pieces in order to establish a scale for each verbal attribute. Using the maximum value for each attribute, informed by the data gathering process, different distance calculations have been implemented to estimate how close an audio file is to the 'best result', represented by the known maximum timbral value. Therefore, by comparing the distance for each attribute value, the one with the shortest distance would be classified as the dominant perceptual quality. However, the results of the different distance measures' testing, detailed in Section 5.4.3, have shown that this approach was not capable of identifying the dominant perceptual quality from a sound composed of different instruments using the calculated timbral values.

Following the unsuccessful results of the initial classification approach, different methods taken from the field of AI research have been investigated in order to develop classification models. Here, machine learning algorithms have been implemented using calculated timbral values as data input. The first method was derived from the unsupervised learning category, which aim to identify classifier functions by performing exploratory data analysis to identify hidden patterns or grouping in sets of data. Here, a k-means algorithm has been performed on a large dataset of timbral values in order to identify five clusters, corresponding to the five verbal attributes, described in Section 5.5.2. While this method has been capable of grouping similar timbral values into five clusters, this method did not provide direct information on

the perceptual qualities due to it using unlabelled data. Thus, a clusters evaluation process is required to identify the grouping, which involves listening and labelling audio files. Testing of the *k*-means method, presented in Section 5.5.2.5, suggested that the proposed clusters' space were representing the five timbral attributes when clusters were relabelled correctly.

The second machine learning approach that has been investigated was supervised learning techniques, where algorithms learn classification models from sets of examples. Thus, the initial task was to create a labelled training dataset to be used to train algorithms. Then, this training corpus, which consisted of 1250 labelled samples as described in Section 5.5.3.2, has been used with two supervised learning algorithms. The first method was a SVM algorithm, described in Section 5.5.3.3, which after a parameter tuning process scored a success classification rate of 0.978. The second method was using ANNs, as detailed in Section 5.5.3.6. The classification model generated by the ANNs algorithm scored a success rate of 0.984, slightly better than the SVM's classification model. From these results, it is evident that supervised learning algorithms are appropriate methods for automatically classifying instrument combinations sounds based on their perceptual qualities informed by specific timbre characteristics. However, these methods involve a prior listening and labelling process in order to train the models. While this process can be seen as an extra task, it can also benefit the training processes.

A method to take advantage of the listening and labelling process has been developed, which was presented in Section 5.5.4. Here, this process is used to reinforce the classification models by putting a weighted value on each training sample. Following the successful classification of the supervised learning approach, a SVM algorithm has been selected, using the weighted values to guide the learning method and put emphasis on the training samples with the highest values. This method also allows users to personalise the training datasets, whether by creating their own training corpus, or by evaluating the 'standard' dataset presented in Section 5.5.3.2, and thus altering the corresponding weight values. This technique offers a solution to overcome the challenges of harnessing perception variations between individuals, which is difficult to manage in computing models.

The research developments presented within this chapter have offered solutions to overcome the limitations of the timbral ranking system by proposing methods to estimate the dominant perceptual quality of instrument mixtures. These methods have been built on the findings presented in Chapter 4, integrating the timbre estimation techniques to calculate timbral values corresponding to the five verbal attributes. From the observations of the results detailed in this

chapter, it is evident that machine learning methods have been able to generate classification models from sets of timbral values, and thus automatically classify sounds by their perceptual qualities, which proposes an answer to **RQ3**. Such methods offer an approach to overcome the need of listening to audio files to manually identify their perceptual qualities, which can also be integrated into generative audio system by enabling an automatic perceptual evaluation of the generated sounds.

5.7 Chapter Summary

This chapter has presented the investigations that have been carried out to develop a method capable of automatically identifying the perceptual quality of a sound created by combining different instruments, using values from the estimations of specific timbre characteristics.

First, the selected verbal attributes and their corresponding calculation methods have been detailed in Section 5.3, which were built on the findings from the timbral ranking system presented in Chapter 4. Then, a classification approach, performing distance measurement onto calculated timbral values, has been explained. However, results of its testing, described in Section 5.4.3, indicated that distance calculations were not appropriate methods for classifying timbral value data.

The second part of this chapter has presented the investigations into using machine learning methods to harness classification models using timbral value data as input. Here, it has been suggested that supervised learning algorithms, where classification models are learned from examples, may be appropriate methods for automatically classifying timbre properties from audio files. This was also supported by their successful classification results. Moreover, a method to take advantage of the listening and labelling process required by supervised algorithms has been developed. Here, the labelling process has been utilised to reinforce the classification models, and also to allow users to personalise the training datasets. Thus, this method proposed a solution to overcome the challenge of perception variations between individuals.

The developments presented in this chapter has highlighted some key outcomes, which are summarised below:

- Definition of a scale for each verbal attribute.
- Distance calculations are not appropriate methods for timbral values classification.
- Supervised learning algorithms present successful classification models.
- Personalisation perception classification models with reinforced supervised learning methods.
- Methods to estimate perceptual quality of a sound resulting from combining different instruments, using its calculated timbral values.

6

Timbral Driven Instrument Combination

6.1 Chapter Overview

This chapter presents investigations into harnessing the findings from developments presented in Chapters 4 and 5 into a computing system designed to generate combinations of audio samples of recorded instrument notes matching specific timbral qualities.

First, the text discusses some of the outcomes of developing a timbral ranking method and a timbral classification system, focusing on their benefits for integrating them into techniques for combining instrumental timbres. The chapter continues with detailing the technical environment of the computing system, which uses two groups of instruments (string and brass) and timbral descriptors as the main criteria for the output decision function. This is followed by discussions about the instrument combination search space and the selected approach to address the different challenges.

The second part of this chapter details the developments for addressing the search space challenges, starting with discussions about the reasons and benefits of incorporating Artificial Intelligence (AI) methods. Then, the text describes and analyses the AI techniques that have been investigated in order to optimise the instrument combination search algorithms. The chapter continues with descriptions of the system's method for input and output information. Finally, the chapter concludes with discussions about the findings and outcomes of developing a system for combining string and brass instrument notes to create specific timbral textures. Such developments have provided insights for expanding identified techniques for controlling

instrument timbre and timbral combination in a computing system processing a larger musical ensemble.

Below is an overview of this chapter's structure:

- 6.2 - Introduction
- 6.3 - Combining String and Brass Sampled Instruments
- 6.4 - Search Space
- 6.5 - Search Algorithm Optimisation
- 6.6 - Sequencing
- 6.7 - Rendering of the Results
- 6.8 - Chapter Conclusions
- 6.9 - Chapter Summary

6.2 Introduction

The developments presented in Chapter 4 have identified methods to calculate timbre characteristics correlating to specific perceptual qualities. Building on these findings, Chapter 5 has suggested techniques to automatically evaluate and classify the perceptual qualities of combined instruments audio files, informed by their calculated timbral values. Such outcomes represent significant methods for harnessing analysis and control of instrumental timbre and timbral combinations.

Following the findings from the previous research developments, it has been decided to investigate the potential of the techniques listed above for controlling the sonic qualities of sounds created by combining audio samples of instrument notes. This resulted in the implementation of a computing system for combining audio samples of instrumental notes. Here, techniques for retrieving timbre characteristics from instrument combination audio files and for automatic timbre classification have been integrated in different aspects of the generative processes. In order to evaluate the feasibility of applying these techniques, the presented system is designed to operate with groups of string and brass instruments. The reason for this choice is to experiment with a smaller instrument combination space and establish techniques for systems that would operate with a larger selection of instruments. Moreover, a string ensemble offers a varied range of sonic possibilities and has been used by several classical and contemporary composers. The group of brass instruments also provides a different palette of sounds from the other group. The structure of the string and brass instruments database is detailed in Section 6.3.2.

The main challenges of generative processes are the search space and the output decision function. In the present system, the generative algorithm has to search across the different string and brass instruments' notes, where space increases exponentially, in order to output a solution that matches different criteria. Section 6.4 details the challenges of the instrument combination search space along with the different available search criteria. The selected approach to address the instrument combination search space, which has utilised AI techniques, is described in Section 6.5.

Sections onwards detail the investigation into developing the generative system for string and brass instrument note combinations. By examining the implementation for a reduced group of instruments, the objective is to establish methods to harness aspects of timbre perception,

using the techniques from the two previously presented systems, for future computing systems designed to operate with a larger selection of instruments.

6.3 Combining String and Brass Sampled Instruments

This section presents the technical environment of the system designed to generate combination of audio samples of string and brass instruments notes that match specific timbral qualities. The text also provides a description of the programming environment and the structure of the instrument database.

6.3.1 Programming Environment

Similarly to the timbre classification system detailed in Chapter 5, the system presented within this chapter is developed on a Macintosh operating system (Mac OS), using the `Python 3.5` programming language. Again, the timbre estimations are processed within the `Matlab`[1] environment, using some functions taken from the `MIRtoolbox 1.6.1`[2]. Following the methods of the timbre classification system, the `Matlab` script is also directly and automatically executed within the `Python` environment.

6.3.2 Instrument Database Organisation

The present system is designed to generate combinations of sampled instrument notes. Thus, it relies on a database of audio files for its search algorithm. Furthermore, the timbre calculations, utilised to evaluate the perceptual qualities of the sound created by the combining instruments, are also performed on audio files, similarly to the other two systems described in the previous chapter. The sound database comprises string and brass instruments' audio samples extracted from the `SOL 0.9 HQ` sound library provided with the *Orchids* program[3]. This sound database, based on the large IRCAM Solo Instruments[4], consists of sounds from 16 instruments, with numerous playing techniques for each. Specific details about the structure of the sound database utilised in the present generative systems are listed in the following sections.

6.3.2.1 Instruments

As stated previously, the generative system has been designed to operate with string and brass instruments. Thus, the sound database comprises four string instruments: bass, cello, viola,

[1] `http://www.mathworks.com/products/matlab/`
[2] MIRtoolbox is available at `https://goo.gl/d61EO0`
[3] `http://forumnet.ircam.fr/product/orchids-en/`
[4] `http://forumnet.ircam.fr/product/ircam-solo-instruments-en/`

and violin. In regards to the brass instruments, there are also four different instruments: French horn, tenor trombone, C trumpet, and bass tuba. For each instrument, three playing techniques have been selected. All the string instruments have similar playing techniques: *note-lasting* (dynamic mezzo-forte *mf*), *pizzicato-secco* (dynamic *mf*), and *staccato*(dynamic *mf*). As its name suggests, the *note-lasting* technique consists of an audio sample of a bowed note's full duration. The *pizzicato-secco* technique consists of audio samples of damped plucked notes, and the *staccato* technique is audio samples of shortened bowed notes.

In regards to the brass instruments, the French horn has the playing techniques *note-lasting* (dynamic *mf*), *staccato* (dynamic *mf*), and *brassy* (dynamic fortissimo *ff*). The tenor trombone has *staccato* (dynamic *mf*), *decrescendo* (dynamic *mf* to pianissimo *pp*), and *sforzando* (dynamic forte *f*). The C trumpet *note-lasting* (dynamic *mf*), *staccato* (dynamic *mf*), and *sforzando* (dynamic *f*). Finally, the bass tuba has *staccato* (dynamic *mf*), *decrescendo* (dynamic *ff* to *pp*), and *sforzando* (dynamic fortepiano *fp*). The *note-lasting* technique consists of a blown note's full duration, the *staccato* technique is a short note, *brassy* consists of an increased lip tension and a strong attack note, *decrescendo* is a note played gradually more softly, and the *sforzando* technique consists of notes with a strong initial attack.

6.3.2.2 Audio Files Specifications

All the audio samples of the sound database are the results of high-quality recordings of individual notes, played with the same intensity. All audio files are uncompressed (encoded in the WAVE audio format). The specifics of the audio files are as follows:

- File type: WAV
- Sample rate: 44.1 kHz
- Bits per sample: 16
- Audio channel: 1 (mono)

6.3.2.3 Sound Database Structure

The structure of the sound database is important for the search exploration algorithm, optimising the search query. Here, each instrument has a folder, which contains a sub-folder for each of the three playing techniques. Then, each playing technique comprises a folder for each

pitch. The audio files represent the different notes for all octaves available for each pitch. Figure 6.1 represents the folders' tree structure of the sound database. This figure focuses on the bass instrument's organisation, which follows the structure $instrument \rightarrow playing\ technique \rightarrow pitch \rightarrow audio\ files$. The number of notes and available octaves for each instrument's playing technique may vary according to the playing range of the instrument. For example, the bass does not have a note above the fourth octave and has 3 to 4 notes for each pitch. Table 6.1 displays the number of different notes for each string instrument's playing technique, and Table 6.2 shows the number of notes for the different brass instruments. It shows that with only two group of four instruments and three playing techniques for each instrument, the sound database consists of numerous notes, representing a significant search space.

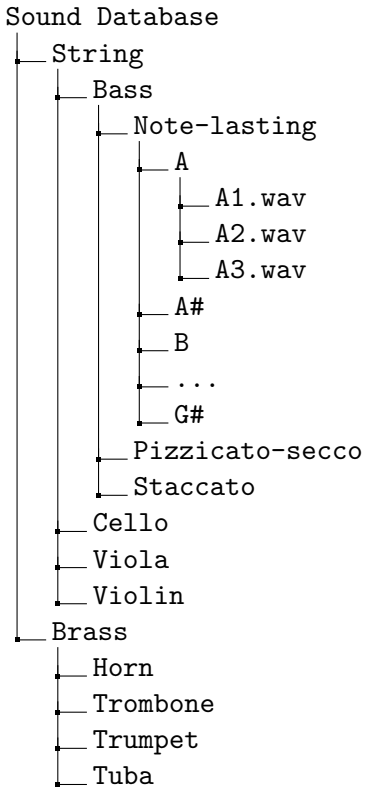

Figure 6.1: Tree structure representation of the sound database, with a focus on the organisation of the bass instrument's audio samples. Each instrument follows a similar structure.

	String Instruments			
Playing Techniques	**Bass**	**Cello**	**Viola**	**Violin**
Note-lasting	36	38	42	38
Pizzicato-secco	37	39	34	32
Staccato	40	38	46	42

Table 6.1: Number of notes for each string instrument's playing technique in the sound database.

	Brass Instruments			
Playing Techniques	**Horn**	**Trombone**	**Trumpet**	**Tuba**
Note-lasting	45	–	29	–
Staccato	41	39	29	33
Brassy	43	–	–	–
Decrescendo	–	28	–	33
Sforzando	–	33	29	33

Table 6.2: Number of notes for each brass instrument's playing technique in the sound database.

6.4 Search Space

This section discusses different aspects of the process for generating combinations of audio samples of instrument notes. Here, the objective is to define the search space of the algorithm designed to output combinations of string or brass instruments' notes matching specific parameters. By defining the search space, this section also outlines the operations of the generative process and describes the type of outputs that can be expected from the present system. This section starts with details about the compositional framework that has been selected for the combination of notes, due to it not being based on a random process. Then, the text details the different criteria that can be defined by the users. These parameters are then utilised to guide the search algorithm for the selection of its outputs.

6.4.1 Instruments Combinations

An important aspect of generative processes is the definition of the decision rules for the outputs. The most basic rule is to output random values. This technique can provide interesting results and has been used for musical purposes by many composers for several decades. However, within this system, the generation of instrument combinations follow a chords framework, which is a compositional technique that consists of playing two or more notes simultaneously. Thus, the timbral combination approach is represented by the generation of chords, where notes can be played simultaneously by one type of instrument, or spread across the different string or brass instruments.

The chord generation rules follow the traditional Western music pitches combination framework. Here, three types of chords have been defined: dyad, triad, and seventh, representing respectively 2, 3, and 4 notes played simultaneously. Two types of dyad have been implemented: major and minor. In regards to the triad and seventh chords, four different types have been selected: major, minor, augmented, and diminished. Tables 6.3, 6.4, and 6.5 show the note combinations for the dyad, triad, and seventh chords respectively. Here, the intervals are calculated from the root note which is defined by the selected key.

In the present investigation, the note combination technique is used to represent the combinations of instruments playing simultaneously. Here, the use of chord combinatorial rules applied on a limited set of instruments is a smaller representation of the instrument combination challenge, which is utilised here to evaluate the feasibility of the approach presented

	Intervals from root
Chords	**Third**
Major dyad	Major
Minor dyad	Minor

Table 6.3: Note intervals for the two different dyad chords rules.

	Intervals from root	
Chords	**Third**	**Fifth**
Major triad	Major	Perfect
Minor triad	Minor	Perfect
Augmented triad	Major	Augmented
Diminished triad	Minor	Diminished

Table 6.4: Note intervals for the four different triad chords rules.

	Intervals from root		
Chords	**Third**	**Fifth**	**Seventh**
Major seventh	Major	Perfect	Major
Minor seventh	Minor	Perfect	Minor
Augmented seventh	Major	Augmented	Major
Diminished seventh	Minor	Diminished	Diminished

Table 6.5: Note intervals for the four different seventh chords rules.

within this study, and thus establishing methods that could be expended to larger musical ensembles. Furthermore, this combinatorial rule is based on harmonic sets. Thus, with only a few basic rules the generated combinations will sound more musical than only outputting random notes played simultaneously. It is also possible to extend the specifications of the instrument combination algorithm by adding other compositional frameworks, and also by defining new combinatorial rules. However, another constraint is to not generate chords outside of the instruments' ranges. Here, a function has been designed to retrieve the different notes that are available in the sound database for each instrument and playing technique. With this function, it is possible to expand the database by adding new instruments and playing techniques, while keeping the chord generations in the available instruments' ranges.

6.4.2 Search Criteria

As explained above, the generative system uses a chord framework for the generation of combinations of instrument notes. Other than the selection of the type of chord, different parameters can also be controlled by the user in order to define their musical ideas. These criteria are also used to specify the type of desired outputs, thus, refining the search query in order not to have a completely random generation. The following sections describe the different criteria that can be defined by the users in order to guide the search algorithm to match their musical ideas.

6.4.2.1 List of Instruments

The first criterion for defining the desired types of output is the arrangement of the musical ensemble. As mentioned in Section 6.3.2.1, the generative system operates with audio samples from four string instruments (bass, cello, viola, and violin) and four brass instruments (horn, trombone, trumpet, and tuba). Thus, the user can define which instruments make up the ensemble. It is also possible to specify the number of instruments for each type.

6.4.2.2 Playing Techniques

Another parameter that is proposed to the user is the selection of the playing techniques. As stated in Section 6.3.2.1, there are three playing techniques for each instrument in the system's database: *note-lasting*, *pizzicato-secco*, and *staccato* for the string instruments, and three techniques from *note-lasting*, *brassy*, *staccato*, *decrescendo*, and *sforzando* for the brass instruments. Thus, the user can define a technique for all the string instruments of the musical

ensemble, for example. It is also possible to assign playing techniques to instruments individually.

6.4.2.3 Perceptual Qualities

The last parameter that can be defined by the user, which is also the most relevant to the study presented in this thesis, is the desired perceptual quality of the instrument note combinations. This parameter is also important in the decision rules for outputting an instrument's combination due to it being the main characteristic of this generative process. Here, same verbal descriptors as those investigated in Chapters 4 and 5 have been selected. In total, five verbal attributes are available: *breathiness*, *brightness*, *dullness*, *roughness*, and *warmth*. The methods for estimating the timbral values corresponding to each of the five verbal attributes are based on the techniques described in Section 5.3.

6.4.3 Discussions

This section aimed to define the search space of the generative process in order to illustrate the challenges of outputting instrument combinations driven by timbre characteristics. As mentioned in Section 6.3.2.1, the present generative system operates only with a group of string or brass instrument audio samples, extracted from the `SOL 0.9 HQ` sound library provided with the *Orchids* program. The use of a smaller group of instruments allows for operating with a smaller search space, while keeping the instrument combination challenges. With this set of groups of four instruments and three different techniques for each, detailed in Tables 6.1 and 6.2, the total of notes combinations possibilities is already significant, highlighting the large search space with only a small set of instruments. Also, string and brass instrument ensembles propose a broad range of musical possibilities, offering a varied sonic palette to the users with a small number of instruments.

While a random process for combining instruments' notes could have produced interesting results, it has been decided to base the note combination function on a chord framework—a compositional technique often used in Western music. Here, sequences of 2, 3, and 4 notes to be played simultaneously have been chosen. These three different sequences are suitable for a group of four types of string or brass instrument, because notes can be spread across the four different types of instruments. The different rules for the chord combinations that have been implemented in the generative system have been detailed in Section 6.4.1. Moreover, by

constructing the notes combinations on harmonic sets, the solutions generated by this system would sound musical according to the Western music standards, while this may not be the case with randomly combined notes. Furthermore, it is possible to modify the notes combination process by altering the chord rules, or by programming different compositional frameworks. With the function designed to retrieve the available notes in the database, the generated chords will not be outside of the instruments' ranges, an issue that can sometimes happen in computer-aided composition systems.

The generation of chords using only string or brass instruments' notes without any other information would result in outputting random combinations, which would not reflect a user's view. Therefore, different parameters can be controlled, which allow the users to define their musical ideas. For instance, it is possible to set the arrangement of instruments, which is an important compositional information, by specifying the type and number of instruments. Section 6.4.2 lists the different criteria that have been implemented in the system. Furthermore, if a parameter is not defined, a parameter's value will be randomly selected from the available range. For example, if no playing technique is specified by the user, the system will randomly select one technique from *note-lasting*, *pizzicato-secco*, and *staccato* for the string instruments, and one of the three available techniques for each brass instrument from *note-lasting*, *brassy*, *staccato*, *decrescendo*, and *sforzando*. This flexibility allows for chance, a creative process appreciated by many composers. These criteria are also important for refining the search space, thus generating sounds that fit the users' musical view.

Once all the parameters are all set, the system can generate chords following the defined criteria. In this initial search algorithm, the list of instruments, playing techniques, and the key are the first parameters to be processed to create a chord according to the rules described in Section 6.4.1. In order to estimate the perceptual quality of the instrument combination, the chord is generated as an audio file, rendered using the audio samples contained in the database. Then, the timbre estimation function is applied in order to calculate the timbral values, which are then input into the timbre classification system detailed in the previous chapter. Once these processes have been performed, the value for the perceptual quality of the chord is obtained, which is then compared with the value selected by the user. If it does not match, a new chord is generated until this criterion is fulfilled. However, it is evident that this technique for evaluating the perceptual quality of the generated chords is lengthy and computationally expensive. Therefore, a method to overcome this long search process is required, the investigation of which is detailed in the following section.

6.5 Search Algorithm Optimisation

This section discusses the different approaches that have been investigated to implement a search algorithm designed to output instrument combinations fulfilling the different criteria, which are set by the users and detailed in Section 6.4. Here, different methods have been tested in order to identify the most relevant technique for evaluating if the note combinations match the desired perceptual quality, which is the main factor in the decision rules. An overview of the note combination search problem, explaining the need for establishing more efficient methods for estimating the perceptual quality of the generated chords, is proposed in the next section.

6.5.1 Problem Overview

Section 6.4 explained that the present computing system is designed to generate note combinations using audio samples from string and brass instruments. Here, a chord combination framework has been selected that combines notes that will be played simultaneously according to the rules detailed in Section 6.4.1. Different parameters can be specified by the users in order to define their musical ideas, which are then processed to refine the search space. From these criteria, discussed in Section 6.4.2, the perceptual quality parameter has an important role in the decision rules, due to it being the main attribute for the generative process of the present system.

As explained in Section 6.4.3, the perceptual quality of a chord is estimated from generating an audio file and performing the timbre estimation function to obtain its timbral values. This information is then input into the timbre classification system, presented in Chapter 5, in order to estimate the chord's perceptual quality. Once this information is retrieved, the generated chord is output only if it matches the user criterion. If it does not, the whole process is reiterated until the generated chord fits the desired perceptual quality. It is evident that processing the perceptual quality evaluations function in this manner is not an efficient technique, because it can take a long period before outputting a chord with the required quality. Therefore, a method needs to be developed in order to bypass the audio generation process and the acoustic analysis to obtain the timbral values of each generated chords.

The developments presented in the previous chapter have shown that some machine learning methods could be applied to create classification models based on calculated timbral values. Following the success of such methods, it has been decided to investigate the potential use of

machine learning algorithms to address the tedious audio generation and timbre estimations processes. While the task of the machine learning algorithms (Chapter 5) was to identify the category (verbal attribute) from calculated timbral values, here, such methods need to create models capable of predicting the timbral values of a set of combined notes. The following sections detail the implementation and testing of different machine learning methods in order to establish a technique to predict automatically the timbral values of generated chords.

6.5.2 Data Acquisition

Similarly to the machine learning approaches that have been investigated for the development of the automatic timbre classification system presented in Chapter 5, the initial task is to create a dataset, which will be utilised to train the machine learning methods. Here, the objective of using machine learning algorithms is to be able to predict the timbral values of a generated chord without having to create an audio file in order to perform the timbre estimations function. Therefore, the dataset needs to contain examples of note combinations with their calculated timbral values, which will serve as training samples.

In order to create the training dataset for establishing predictive models, numerous chord combinations have been generated using groups of string instruments and brass instruments, following the rules detailed in Section 6.4.1. Here, 10,000 chords for each type of note combination have been created, with the selection of instruments and playing techniques being performed randomly. Then, each chord has been rendered as an audio file, which resulted in creating 30,000 audio files for groups of string instruments and 30,000 audio files for groups of brass instruments. The timbral values for each audio file have been calculated using the timbre estimations function presented in Section 5.3. Thus, the training dataset consisted of 30,000 samples of note combinations with their corresponding timbral values for each instrument group. Entries of the dataset are constructed as follows: [Instruments], [Playing techniques], [Pitches] = [$tBreathiness$, $tBrightness$, $tDullness$, $tRoughness$, $tWarmth$]. Once this training dataset is created, it is possible to apply different machine learning methods to identify predictive models, which are described in the next sections.

6.5.3 Regression Models

The developments presented in Chapter 5 have shown that machine learning methods can be applied to classifying audio files according to their timbre characteristics. Such methods have

been able to successfully learn classification models from the analysis of training datasets. However, here, the required task for such functions is to predict data based on the input information. Thus, instead of using machine learning to create classification models, designed to identify the category that input values belong to, here such algorithms are utilised to create regression models, which aim to predict the values from a set of variables. The objective of using regression methods is to be able to predict the timbral values directly from the information about the instruments, playing techniques, and pitches. Therefore, it would be possible to omit the audio generation and timbre estimation processes, mentioned in Section 6.4.3, in order to retrieve the generated chords' timbral values utilised to evaluate their perceptual quality. Such methods would speed up the process of comparing the chord's quality with the user's selection. The following sections detail the different machine learning algorithms for regression models that have been investigated in order to predict timbral values from sets of instrument information.

6.5.3.1 Support Vector Machines (SVM) for Regression

Following the successful results of using a Support Vector Machines (SVM) algorithm for the classification task (Section 5.5.3) the first method for learning regression models is based on this algorithm. Here, instead of trying to find a delimited space for each category, with a gap that is as wide as possible between the different categories, the SVM learns the regression function from analysing the samples of the dataset, which is then used for predicting values. Figure 6.2 presents a theoretical example to illustrate the regression problem formulation. Here, three continuous-valued functions have been generated using three different kernels with the objective to approximate the data points with minimum error margins.

The SVM algorithm has been implemented in `Python 3.5` using the `svm.SVR` function, which is a SVM method for regression task, called Support Vector Regression (SVR), taken from the `Scikit-Learn v0.18` package. Here, the training dataset, detailed in Section 6.5.2, has been divided by samples from each type of chord (dyad, triad, and seventh) and each type of instrument (string and brass). The input data consisted of vectors of the information about the note combinations encoded as numbers and vectors of their corresponding calculated timbral values as output data. Then, the `svm.SVR` function has been applied to each type of chord and instrument. This process resulted in creating six regression models, one for each of the three chord combinations and of the two groups of instruments.

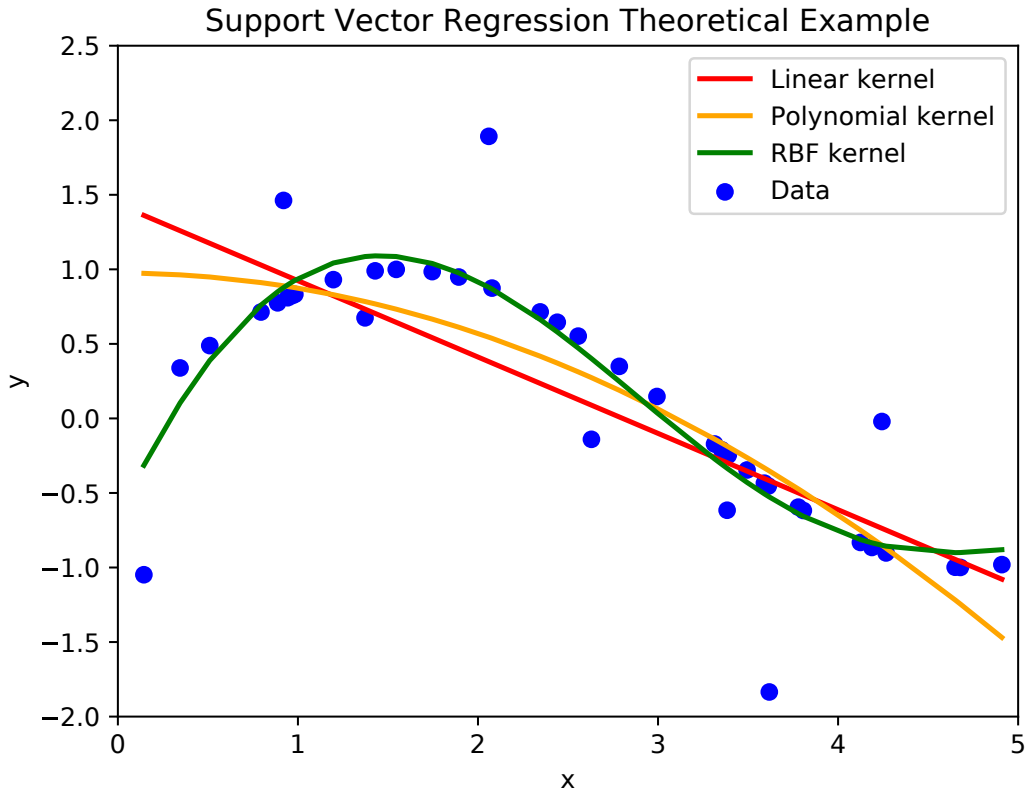

Figure 6.2: Theoretical example of one-dimensional regression models using linear, polynomial, and RBF kernels based on randomly generated values.

Similarly to the implementation of supervised learning methods detailed in Section 5.5.3, a parameter tuning process has been performed to obtain the best parameter combination for the `svm.SVR` function for each of the three types of chords. Here, the `Scikit-Learn`'s `model_selection.GridSearchCV` function has been utilised to test and compare the performances of different combinations of parameters. Table 6.6 lists the `svm.SVR` function's parameters for each of the six regression models with their corresponding calculated coefficient of determination R^2, providing information about how a predictive model fits the data, produced with 10 000 samples (90% training–10% testing). The coefficient of determination R^2 is calculated as follows:

$$R^2 = 1 - \frac{SS_{reg}}{SS_{tot}} \tag{6.1}$$

where SS_{reg} is the sum-of-squares from the model and SS_{tot} is the sum-of-squares from the

Regression Model	Kernel Type	Kernel Coefficient γ	Penalty Parameter C	R^2 Score
String Dyad	*RBF*	0.1	100	0.345
String Triad	*RBF*	0.1	100	0.336
String Seventh	*RBF*	0.1	100	0.351
Brass Dyad	*RBF*	0.1	100	0.715
Brass Triad	*RBF*	0.1	100	0.566
Brass Seventh	*RBF*	0.1	10	0.453

Table 6.6: Best `svm.SVR` function's parameters for each regression model with corresponding coefficient of determination R^2 produced with 10 000 samples (90% training–10% testing).

horizontal line. The best possible R^2 score is 1.0. It can also be negative if the model does not follow the trend of the data and fits worse than a horizontal line. From the results shown in Table 6.6, we can note that a SVM with a RBF kernel has produced the best models for all types of chords. Furthermore, the best prediction scores have been obtained using the same RBF kernel coefficient γ and penalty parameter C for all, except for the *Brass Seventh* model where penalty parameter C differs from the others.

6.5.3.2 Artificial Neural Networks (ANNs) for Regression

The second machine learning algorithm for regression analysis that has been implemented is using Artificial Neural Networks (ANNs). Building on the successful classification models suggested by this technique, as detailed in Section 5.5.3, an ANNs algorithm has been tested to learn regression models from the dataset discussed in Section 6.5.2. Here, the aim of this machine learning technique is to establish the continued-values function $f : X \rightarrow Y$, where X is all the input data points x_i representing the instrument combinations' information and Y is all the output values y_i representing the calculated timbral values.

The ANNs algorithm uses the `neural_network.MLPRegressor` function, which is a multi-layer perceptron regressor method, also taken from the `Scikit-Learn v0.18`. Similar to the SVM implementation presented in the previous section, the `neural_network.MLPRegressor` function has been applied onto each type of chord from the training dataset using each group of instruments, thus creating two regression models for dyad, triad, and seventh chords. An identical data input procedure to the one utilised in the previous method (Section 6.5.3.1) has been followed. A parameter tuning process has also been performed for each regression

Regression Model	Activation	Solver	R^2 Score
String Dyad	*Relu*	*LBFGS*	0.481
String Triad	*Relu*	*LBFGS*	0.389
String Seventh	*Relu*	*LBFGS*	0.316
Brass Dyad	*Relu*	*LBFGS*	0.764
Brass Triad	*Relu*	*Adam*	0.617
Brass Seventh	*Relu*	*Adam*	0.527

Table 6.7: Best `neural_network.MLPRegressor` function's parameters for each regression model with corresponding coefficient of determination R^2 produced with 10 000 samples (90% training–10% testing).

model of the three types of chords to obtain the best ANNs parameter combination. Again, the `Scikit-Learn`'s `model_selection.GridSearchCV` function has been utilised to test and compare the performances of different combinations of `neural_network.MLPRegressor` function parameters. Table 6.7 lists the `neural_network.MLPRegressor` function's parameters for each of the six regression models with their corresponding coefficient of determination R^2, produced with 10 000 samples (90% training–10% testing). Here, the best regressions models have been obtained with the rectified linear unit (ReLU) activation function, similarly to the ANNs classification models (Section 5.5.3.6). For the solver parameters, two different optimisation algorithms have performed best. For the *Brass Triad* and *Brass Seventh*, the stochastic gradient-based optimiser *adam* (short for Adaptive Moment Estimation) produced the best regression models. For the other models, the solver L-BFGS (short for Limited-memory Broyden-Fletcher-Goldfarb-Shanno) has proposed better continued-values functions. Here, the aim of this optimiser is to minimise the function $f(x)$, where x is a vector and f is a differentiable scalar function [262].

6.5.4 Discussions

This section has presented the implementation of two machine learning algorithms designed to predict timbral values from a set of instrument information to overcome the audio generation and timbre estimation processes. The training dataset, detailed in Section 6.5.2, has been created by randomly generating 10,000 chord combinations for each of the three types and the two groups of instruments. Each chord has been generated as an audio file, using audio

samples from the sound database, on which timbre estimations have been performed to calculate its timbral values. This process resulted in the creation of a training dataset composed of 30,000 samples of string note combinations and 30,000 samples of brass, associated with their calculated timbral values, and grouped by types of chord.

The two machine learning algorithms have been applied onto the training dataset to identify a regression model for each type of chord, and thus be able to predict timbral values from the information about the instruments, playing techniques, and pitches that compose the chords. Following the success of the supervised learning techniques utilised for the classification task, and discussed in Section 5.5.3, similar techniques have been adopted and implemented, but for a regression task. The coefficient of determination R^2 of each regression model using SVM and ANNs algorithms can be found in Tables 6.6 and 6.7, respectively. These scores suggest that the ANNs algorithm identified more accurate regression models for all the brass and string chords, except for string's seventh chords, where the SVM algorithm produced a slightly higher score. These prediction scores are not as high as the classification scores from Chapter 5. Nevertheless, a perfect prediction of the timbral values is not fundamental here, especially with values using a 4 number float, but rather a close estimation, which will be used for identifying the correspondent perceptual quality. Moreover, the audio generation and timbre estimation processes can still be performed in order to calculate the true timbral values from the audio file. An increase of training samples could result in an improvement of the regression model's performance, which is suggested by the learning curves of the SVM models shown in Figures 6.3 and 6.4, and in Figures 6.5 and 6.6 for the ANNs regression models. Nevertheless, these performances suggest that the ANNs and SVM algorithms are capable of producing regression models for predicting timbral values of chords generated by the system, without having to generate an audio file and perform the timbre estimations process.

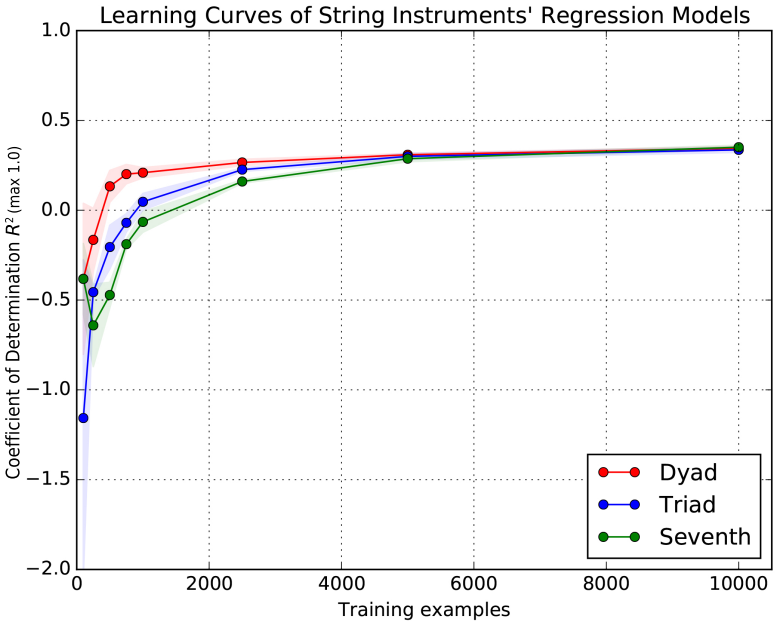

Figure 6.3: Learning curves of the string instruments' regression models produced by the SVM algorithms.

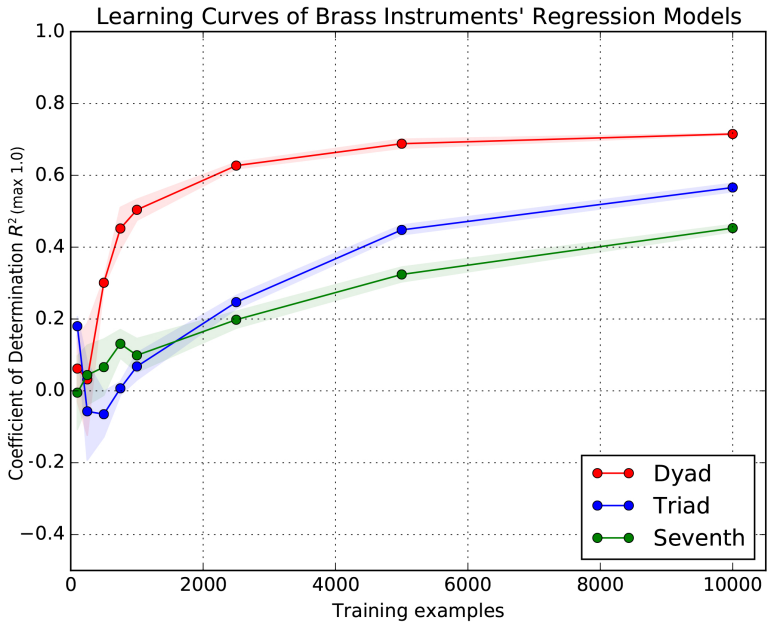

Figure 6.4: Learning curves of the brass instruments' regression models produced by the SVM algorithms.

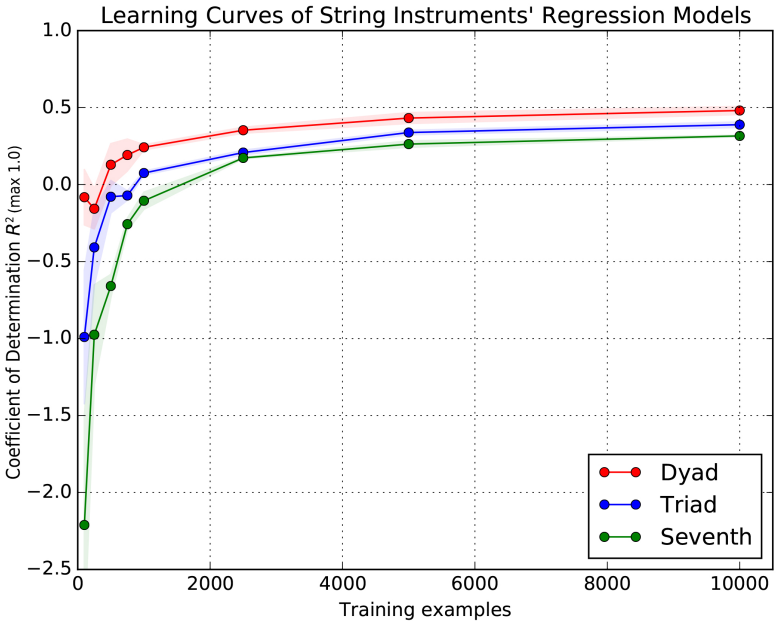

Figure 6.5: Learning curves of the string instruments' regression models produced by the ANNs algorithms.

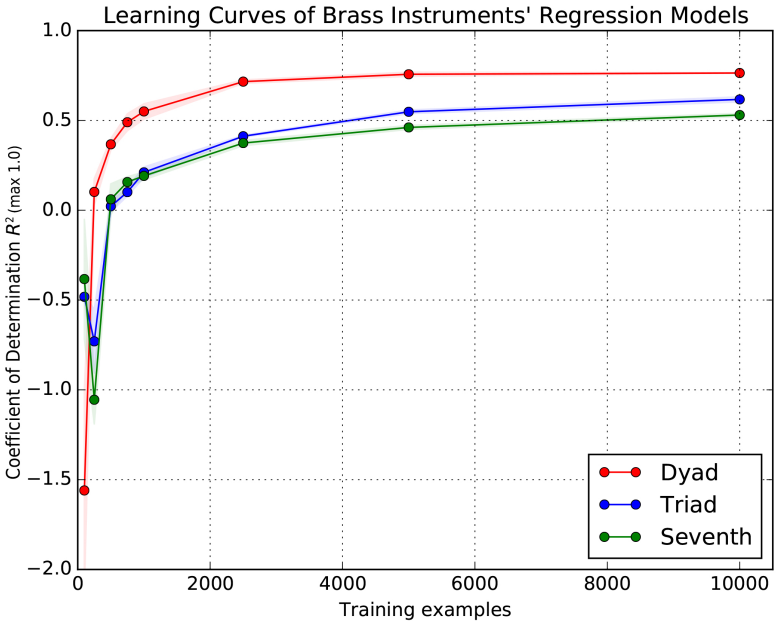

Figure 6.6: Learning curves of the brass instruments' regression models produced by the ANNs algorithms.

6.6 Sequencing

This section reviews the different techniques that have been implemented to define the desired perceptual quality, and also the number of generations of instrument combinations. This process enables the users to construct a set of different combinations that will then be generated sequentially by the system. The two different input methods are described in the following sections.

6.6.1 Text Input

The first method to control the output of the generative system is by using a text input. Here, the user is able to directly specify the desired perceptual quality for each generation, by manually entering *breathiness*, *brightness*, *dullness*, *roughness*, or *warmth*. Then, the list of verbal attributes is used as parameter for the generation of instrument combinations, following the order of the written verbal attributes. The number of verbal attributes contained in the list is also utilised to define the number of combinations that will be generated. For example, if the list of verbal attributes is [*dullness*, *brightness*, *brightness*, *roughness*], the system will generate four chords, each matching their corresponding perceptual quality in the list.

6.6.2 Audio File Input

The second method for defining the sequence of generations is based on the input of an audio file. This method enables the users to input an audio file, which will be used as a model for the creation of the sequences of combinations. However, here, the system does not try to match the instrument and note content of the musical pieces, which has been the objective of other computer-aided composition systems (Section 2.5). Rather, the generation of sequences is based on the evolution of the perceptual qualities throughout the musical piece. The details of this method for defining and sequencing the output of the generative system are discussed in the following sections.

6.6.2.1 Audio Split

The first step is to split the musical pieces containing traditional Western instruments into shorter audio files, as it would not be possible to estimate timbre properties from the complete audio file. It would also miss the perceptual qualities' evolution, and thus, not create a sequence of attributes.

This method accepts only pieces as an uncompressed WAVE audio file. Once the audio file is selected, the function slices the input at every n seconds interval, number being defined by the user. By default, the piece is split into 4 seconds audio files, following the methods utilised in Chapters 4 and 5. Longer audio file duration is possible, however, the timbre estimation method may not perform accurately. Moreover, the short audio files are generated with a 20 ms fade in and fade out, in order to avoid clipping, which could alter the spectral information of the sound, and thus would return inaccurate timbre estimations. After the splitting process is complete, and all the new audio files have been generated, the timbre estimation function can be performed on the short audio files, in order to evaluate their perceptual qualities.

6.6.2.2 Timbre Analysis

Once the splitting process is complete, the newly generated audio files can be analysed to evaluate their perceptual quality. Here, different timbre properties, corresponding to the five verbal attributes utilised throughout this study, are calculated following the methods presented in Section 5.3. The calculated timbre values are then processed into the timbre classification system, detailed in Chapter 5, in order to retrieve their correspondent verbal attribute. The audio files being analysed following the timeline of the original audio file, the retrieved verbal attributes from the short audio files are listed in the same order. The structure represents the evolution of the perceptual qualities of the musical piece. The sequence of verbal attributes can be further manipulated if desired.

6.6.2.3 Creation of New Sequences

The users are able to use the sequence of verbal attributes retrieved from an audio file to specify the type and number of instrument combinations to be generated by the system. They are also able to manipulate that sequence, enabling them to alter the original evolution of the perceptual qualities. Here, a Markov chain algorithm has been implemented, with its orders needing to be defined by the user. This stochastic process [223], which have been applied in several musical applications as mentioned in Section 3.4.3.1, bases its data generation on its input data, which can be controlled by its orders. Thus, by selecting low-n orders, the Markov chain algorithm will generate sequences of verbal attributes more or less distant to the original sequence, while high-n orders will produce sequences reminiscent to the original set of attributes. Moreover, higher orders need more data to train. This process enables the users to

control the similarity of the generation of sequences of verbal attributes based on the evolution of the original piece, which will then be used as input information for the generative processes. Methods for manipulating the sequences of verbal attributes can be expanded by implementing different algorithms and rules.

6.7 Rendering of the Results

The instrument combinations matching specific perceptual qualities suggested by the system are output as an audio file. This is possible with the use of the audio samples from the sound database, described in Section 6.3.2. Here, a function has been designed to retrieve the audio samples of the instrument notes corresponding to the chords generated by the system. This is facilitated by the consistent formatting of the chord information, which follows the order $instrument \rightarrow playing\ technique \rightarrow pitch$. This formatting pattern has been adopted from the structure of the sound database, described in Section 6.3.2.3. Thus, the function starts by looking at the instrument's name, then the playing technique, and finally the pitch to select the audio sample. The different notes that compose the chords are then combined together, and rendered as uncompressed audio files, encoded in the WAVE audio format, in order to keep all the sonic information. The audio files are generated following the specifics of the sound database files, which are detailed below:

- File type: WAV
- Sample rate: 44.1 kHz
- Bits per sample: 16
- Audio channel: 1 (mono)

Furthermore, the chord audio files are labelled following the order of the input sequence. This enables the users to listen to the audio files chronologically, representing the evolution of the perceptual quality as defined by the input information. Finally, details of each generated combination (instruments, playing techniques, and notes) are displayed, which information can then be further processed by the users.

6.8 Chapter Conclusions

The research developments presented in Chapters 4 and 5 have proposed a method for calculating timbre characteristics, representing perceptual qualities, from audio files of instrument combinations, and a technique for automatic classification of sounds created by combining instrument timbres according to their perceptual qualities. This chapter has reviewed the investigations that have been developed to harness these research outcomes into a computing system designed to combine audio samples of instrument notes matching specific perceptual qualities.

The study for harnessing some of the methods put forward by the research discussed in the previous chapters has been conducted throughout the development of a generative system for string and brass instruments. The rationale for using only groups of string and brass instruments was to reduce the instrument combination space, while still keeping the problem of combining several instruments. Using two groups of different types of instruments also allows the system to operate with different types of sound and outputs. Furthermore, string instrument ensembles offer a broad range of musical possibilities with only a small number of instruments, which have been used by many classical and contemporary composers. The objective of this system was to generate combinations of notes that will be played simultaneously, which could be an approach that encompasses the phenomenon of timbre blending. Here, the idea was to distribute the notes across the different instruments, which could also be done similarly for a larger musical ensemble. Thus, the techniques developed within this generative system for a small group of string or brass instruments could then be applied to systems operating with larger instrument ensembles.

The generative system presented in this chapter operates with audio files for its generative process. Following the techniques utilised in the two timbral systems presented in the previous chapters, the evaluation of the perceptual qualities is performed on the generated audio files. Therefore, a sound database has been created, which consists of high-quality audio recordings of notes from four string instruments (bass, cello, viola, and violin) and four brass instruments (French horn, tenor trombone, C trumpet, and bass tuba). Each instrument also has three different playing techniques: *note-lasting*, *pizzicato-secco*, and *staccato* for each string instrument, and three techniques among *note-lasting*, *brassy*, *staccato*, *decrescendo*, and *sforzando* for each brass instrument. In total, the sound database is composed of 462 different string instruments' notes, as detailed in Table 6.1, and 415 for the brass instruments, as detailed in Table 6.2. A full description of the content and structure of the sound database can be found in Section 6.3.2.

The number of different notes comprised in the sound database mentioned above is already a significant search space. Applying a random generative process on this search space would result in notes being arbitrarily combined, and not necessary musically interesting. Therefore, the rules for combining instrument notes have been based on the chord framework, a compositional technique often utilised in Western music. Here, three types of chords have been implemented: dyad, triad, and seventh. Different variants of these chords have been selected, as described in Section 6.4.1. Furthermore, these combinatorial rules are based on harmonic sets, thus, with only a few basic rules, the generated combinations would sound more musical than a combination of notes randomly generated. It is also possible to extend the instrument combination techniques by modifying the rules of the chords, or by adding different compositional frameworks. Moreover, different criteria have been integrated, which can be controlled by the users in order to define their musical ideas. Here, the users are able to specify the different instruments, playing techniques, and the desired perceptual quality for each chord, which will be processed by the system to refine the search space in order to output chord combinations that match the user's musical ideas. The generative process being driven by timbre properties, the selection of the desired perceptual quality is the main variable in the decision rules for outputting a chord generated by the system. This involves calculating the timbral values of each generated chord by creating an audio file, performing the timbre estimations function, and then inputting the timbral values into the timbre classification system, in order to evaluate its perceptual qualities and being able to compare it with the attribute defined by the users. It is evident that this method was tedious and computationally expensive.

Following the success of the machine learning methods for classification tasks presented in Chapter 5, two machine learning algorithms have been investigated in order to address the audio generation and timbre estimation processes. These techniques have been used to create regression models for predicting timbral values directly from a set of information about the instruments, playing techniques, and notes that compose the generated chord. In order to train the machine learning algorithms, a training dataset has been created by randomly generating 10,000 chord combinations for each of the three types of chords and the two groups of instruments. These combinations have been generated as audio files in order to perform the timbre estimation function to calculate their timbral values. Thus, the training dataset consisted of 30,000 samples of string instrument note combinations, and 30,000 samples using brass instruments, associated with their calculated timbral values, which have been grouped by types of chords, as detailed in Section 6.5.2. SVM and ANNs algorithms have been applied to the

training dataset to create two regression models for each of the three types of chords. Here, the ANNs algorithm has proposed the best regression models, as suggested by the results presented in Section 6.5, and thus, offering a method for predicting timbral values directly from the instruments' information of the generated chords. These regression models allow a quicker process for retrieving the chords' timbral values, which accelerates the comparisons between the chords' perceptual qualities and the attributes selected by the users.

The last part of this chapter has discussed the functions designed for handling the input and output information of the generative instrument combination system. As mentioned previously, different criteria can be controlled by the users, with the perceptual quality parameter being the fundamental variable in the decision rules for outputting the chords generated by the system. Two methods for sequencing the desired perceptual qualities have been implemented. The first method involves writing the sequence of verbal attributes directly, while the second uses an audio file to create a sequence. For the second method, the audio file is processed in order to retrieve the evolution of perceptual qualities throughout the musical piece, following the techniques detailed in Section 6.6, and thus creating a sequence of attributes based on the piece. A Markov chain algorithm has also been implemented, enabling the users to manipulate the sequences of verbal attributes obtained from the analysis of the audio files, creating new sequences with different levels of reminiscence. In regards to the outputs of the system, the generated chords can be rendered as an audio file, enabling the users to evaluate sonically the instrument note combinations simulated by the system. Information about instruments, playing techniques, and notes are also displayed.

The research developments presented within this chapter have proposed methods for incorporating different timbre properties into a computing system designed to generate combinations of instruments' notes, using techniques suggested by the systems discussed in Chapters 4 and 5. This generative instrument combination system for string and brass has offered an answer to **RQ4** by using timbre properties. Such information is utilised to represent perceptual qualities and is the fundamental criterion for a generative process designed to output combinations of notes to be played simultaneously by several instruments. Such methods harnessing timbre properties could be applied into other computing systems that process musical instrument combination. Here, the processes for combining instruments are based on selecting adjectives representing perceptual qualities the instrumental mixtures would produce, and thus, matching combinations of notes with types of sound, or textures, instead of trying to match the content of a target sound, as used in other systems (Section 2.5). The use of AI methods has enabled

the processing and manipulation of timbre properties, underlying perceptual qualities of sound, and the prediction of timbral values from instrument information, thus, bypassing the need to perform acoustic properties calculation on audio sources in order to retrieve timbral information. The methods developed for the present generative system put timbre and its perception as the central element for addressing the challenge of combining instruments and notes.

6.9 Chapter Summary

This chapter has presented the investigations that have been carried out to harness timbre characteristics into computing systems, which has been done throughout the development of a generative system for sampled notes of string and brass instruments.

First, the text has provided information about the programming environment and the structure of the instrument database that has been utilised. Here, audio samples from four types of string instruments and four types of brass instruments, with three playing techniques for each, have been selected. The chapter continued with discussions about the search space for combining instruments' notes and the different parameters that can be defined by the users, details of which can be found in Section 6.4. Here, the note combinations are based on a chord framework, which can be expanded by adding other chord rules or compositional frameworks. Nevertheless, the initial method for estimating the perceptual quality of the generated chords, which involves an audio generation and timbre estimation process, happened to be tedious and computationally expensive.

The second part of this chapter has presented the investigations into using machine learning algorithms to bypass the need for the audio generation and acoustic analysis of each generated chord. Following the success of the supervised learning methods for timbre classification, discussed in Section 5.5.3, similar techniques have been implemented for identifying a regression model for each of the three types of chords. The SVM and ANNs algorithms have been applied to a training dataset composed of 30,000 randomly generated string chords and 30,000 randomly generated brass chords with their corresponding calculated timbral values. Here, the regression models generated by the ANNs algorithms has performed the best prediction scores, suggesting that this technique is able to predict the timbral values of a chord directly from the information about its note combination. Thus, this method has been a solution to address the need of the computationally expensive audio generation and timbre estimations processes.

The chapter continued with discussions about the two methods that have been designed for the sequencing of the instrument combinations. Here, the users can define the desired perceptual qualities by writing the verbal attributes. They can also import an audio file containing traditional Western instrument, which is processed by the system in order to retrieve its perceptual qualities throughout the musical piece. This creates a sequence of verbal attributes which can be used as input information for the generative process. A Markov chain algorithm has been implementing, which enables the users to create new sequences of verbal attributes based

on the evolution of perceptual qualities of the audio file. Finally, the text has discussed the function for rendering the combinations of instrument notes generated by the present system as an audio file. Information about utilised instruments, playing techniques, and notes are also displayed, which can be used by the users for further processing.

The investigations for harnessing the findings from the research developments discussed in Chapters 4 and 5 into a generative instrument combination system has highlighted some key outcomes, which are summarised below:

- Application of timbre estimation and automatic timbre classification techniques into a computing system for combining sampled instrument notes.
- Use of timbre properties to drive generative instrument combinations process.
- Methods for predicting timbral values from instrument information of generated chords.

7

Generated Instrument Combination Examples

7.1 Chapter Overview

This chapter presents a selection of examples to demonstrate the functioning and abilities of the timbral driven generative instrument combination system discussed in Chapter 6. These examples illustrate the types of outputs that can be expected by such a system designed to generate combinations of sampled instrument notes, using two different types of inputs (text and audio).

After defining the scope of this chapter and the examples selected to illustrate the capabilities of the techniques developed throughout this study, the text starts by presenting combinations of instrument generated following text input information. Here, one example, either using groups of string or brass instruments, are described and discussed for each of the five implemented timbral descriptors. The second part of the chapter details sequences of instrument combinations generated using different input information, in order to illustrate the sequencing possibilities. The chapter then concludes with discussions and reflections on the various sets of examples presented throughout the text.

The structure of this chapter is as follows:

- 7.2 - Introduction
- 7.3 - Examples of Instrument Combinations
- 7.4 - Sequencing Combinations of Instrumental Timbres

- 7.5 - Chapter Conclusions
- 7.6 - Chapter Summary

7.2 Introduction

The research findings presented in Chapters 4, 5, and 6 have led to the development of a computing system capable of generating combinations of sampled instrument notes matching specific perceptual characteristics. A selection of examples are presented within this chapter, in order to illustrate the abilities of the generative system.

This chapter intends to highlight the types of instrument combinations that can be generated using the timbral driven system detailed in the previous chapter. The selected approach for developing the system is not to generate complete musical sequences. Rather, the objective is to establish techniques and framework for systems to generate musical materials to be further processed by the users. It also aims to provide techniques for the analysis and control of timbral qualities emerging from combining instrument timbres.Thus, the objective of this chapter is to demonstrate the abilities of the techniques and systems that have been established for analysing and controlling instrumental timbre combinations.

It is important to mention that all the examples described in this chapter do not aim to present any musical characteristics except the note combination techniques being based on Western music chords framework. These short examples are only shown to illustrate the types of outputs generated by the system and demonstrate the functioning of the computing system. The examples have been selected from an engineering point-of-view to highlight the compliance of the outputs to the input parameters and not from a musically trained person approach.

Sections onwards detail various combinations of audio samples of instrument notes created by the generative system, using different types of input information and ensembles. Sampled instrument notes have been taken from the audio files database detailed in Section 6.3.2. Here, the audio samples are all labelled using the same format: $instrument \rightarrow playing\ technique \rightarrow note$. This labelling is utilised to retrieve the musical information of the sample from its filename. It is also essential in the process for selecting the audio files to combine. Each example presented in this chapter has been rendered as an audio file, where instrument combinations output by the system have been concatenated without any further audio processing. All examples are available online at the following address: `http://www.aurelien-antoine.fr/thesis_audio/`.

7.3 Examples of Instrument Combinations

This section presents a selection of instrument combination examples generated by the timbral driven generative system detailed in Chapter 6. One example for each of the five implemented timbral descriptors (*Breathiness, Brightness, Dullness, Roughness,* and *Warmth*) using groups of string or brass instruments are considered.

Example 1 - Breathiness The first example has been generated using the timbral descriptor *Breathiness* as a target for combining 2 brass instruments from the ensemble horn, trombone, trumpet, and tuba. The combination of instrument notes presenting qualities of the attribute *Breathiness* has been generated by the system with the following information (corresponding to the audio file labelling format and displayed here as instrument, playing technique, note):

Trumpet, sforzando, C♯4

Tuba, decrescendo, F3

Figure 7.1 shows a spectrogram of the audio file generated by combining the sampled instrument notes detailed above. Here, we can note noise in the high frequencies (the yellow parts above 6 kHz), a characteristic related to the attribute *Breathiness* (Section 4.5).

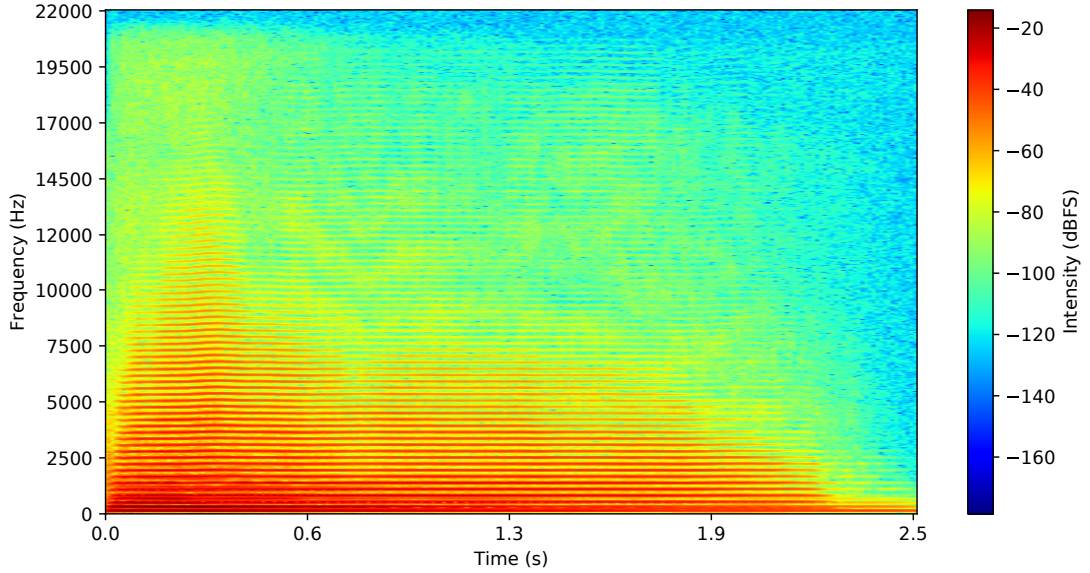

Figure 7.1: Spectrogram of the audio file generated for the instrument combination of *Example 1 - Breathiness*.

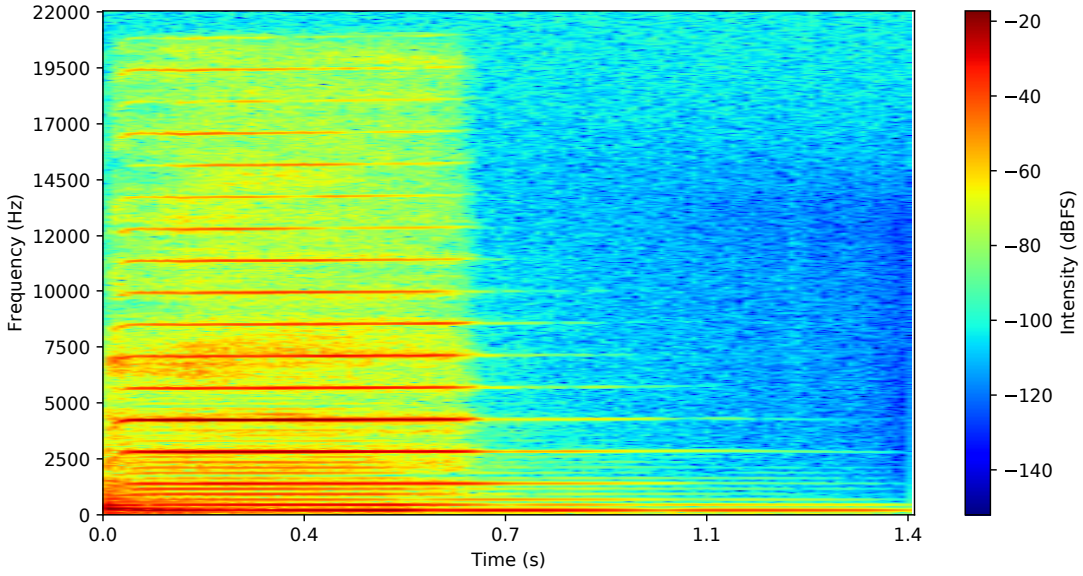

Figure 7.2: Spectrogram of the audio file generated for the instrument combination of *Example 2 - Brightness*.

Example 2 - Brightness The second example presents an instrument combination for the timbral descriptor *Brightness*. Here, 3 string instruments have been combined by the system, with the following details for each sample:

Violin, note-lasting, F6

Viola, pizzicato-secco, D3

Bass, note-lasting, A♯3

A spectrogram of this instrument combination is presented in Figure 7.2. It shows that this audio file contains a large amount of energy in the high frequencies, a feature of the *Brightness* quality, and demonstrated by the orange and red parts at the top.

Example 3 - Dullness For the timbral descriptor *Dullness*, the combination has been generated with 2 string instruments, selected from the ensemble bass, cello, viola, and violin. The suggested notes were as follows:

Cello, pizzicato-secco, D♯3

Bass, pizzicato-secco, G2

Figure 7.3shows a spectrogram of this combination, where we can denote a low amount of high frequencies represented by the presence of green and blue colours above 2.5 kHz.

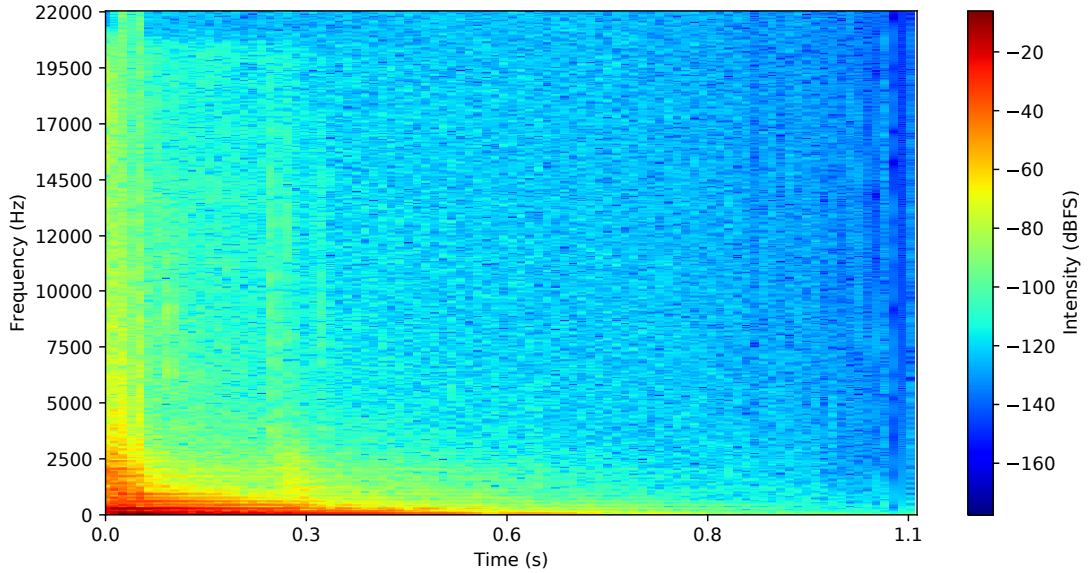

Figure 7.3: Spectrogram of the audio file generated for the instrument combination of *Example 3 - Dullness*.

Example 4 - Roughness The instrument combination for the attribute *Roughness* has been generated with 5 brass instruments. The following sampled notes were combined:

Trumpet, sforzando, G♯5

Trumpet, note-lasting, F♯4

Tuba, sforzando, C♯3

Trombone, sforzando, F2

Horn, note-lasting, A♯2

A spectrogram of the audio file generated for this instrument combination is shown in Figure 7.4. Here, we can note the presence of high energy in different frequency bandwidths, represented by the orange and red horizontal lines in the figure. Furthermore, distances between bandwidths are relatively short, a characteristic of the sensation of roughness.

Example 5 - Warmth Finally, the combination for the timbral descriptor *Warmth* as the target has been generated with 3 string instruments selected from the ensemble bass, cello, viola, and violin. The information about each sample was as follows:

Violin, pizzicato-secco, F♯4

Cello, pizzicato-secco, B2

Bass, staccato, D♯2

Figure 7.5 presents a spectrogram of this instrument combination. It shows that the energy is mainly in the low frequencies, characterised by the orange and red colours below 2500 Hz.

This section has presented an example of instrument combinations for each of the five implemented timbral descriptors. Here, the information about the sets of instruments to combine and the target descriptors was manually input in the generative system. Such examples have illustrated the methods for generating individual instrument combinations matching defined timbral qualities. Furthermore, the use of different instrument types and ensembles have demonstrated the abilities of the system to generate combinations involving a wide range of sonic possibilities. These examples also depict the types of outputs that can be expected using this sampled instrument note combination system.

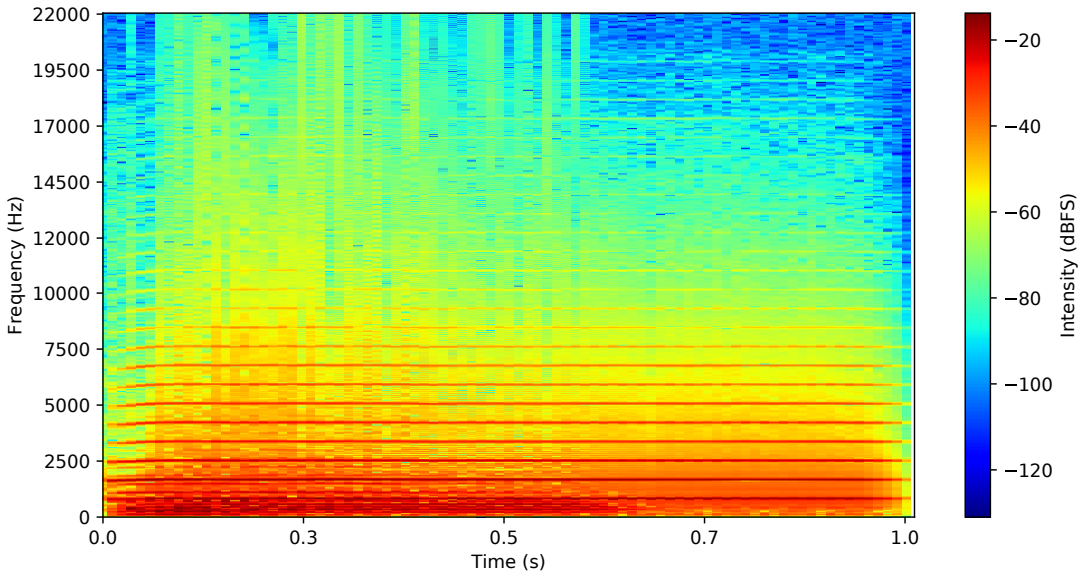

Figure 7.4: Spectrogram of the audio file generated for the instrument combination of *Example 4 - Roughness*.

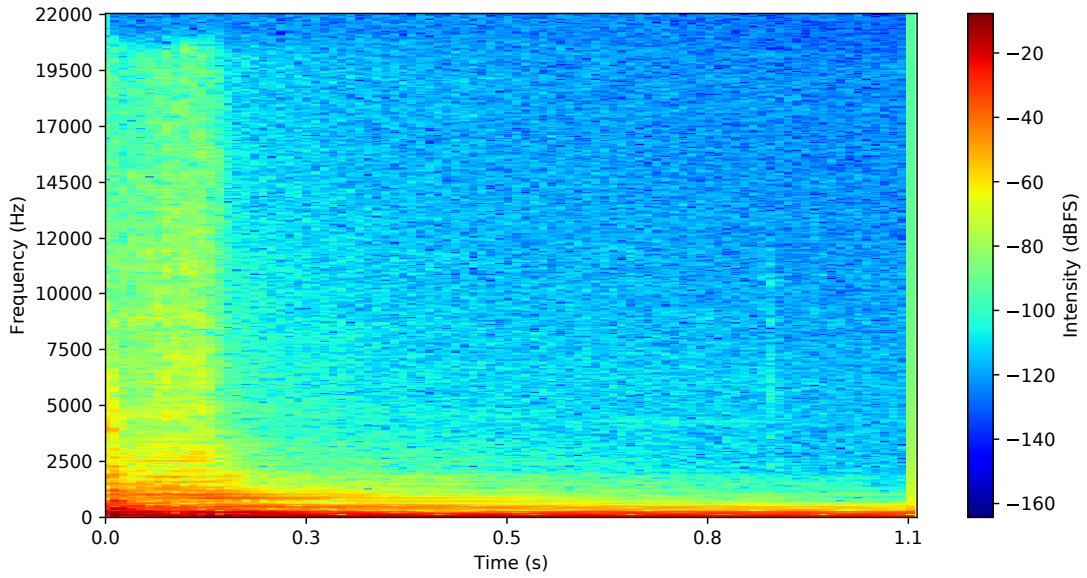

Figure 7.5: Spectrogram of the audio file generated for the instrument combination of *Example 5 - Warmth*.

7.4 Sequencing Combinations of Instrumental Timbres

The second part of this example chapter presents a selection of sequences of instrument combinations. Three examples using different sequencing criteria are discussed in order to highlight the potential of the system's combination abilities.

Example 1 - From audio file analysis The first example has been generated from processing an audio file. The objective of using this method is to produce sequences that follow the timbral structure of the file. Here, the timbre estimations have been performed on an arbitrarily selected 8 seconds excerpt of Beethoven's *Symphony #5 - Allegro Con Brio*. This process produced a sequence of four timbral descriptors: *Dullness, Brightness, Roughness*, and *Brightness*. This set has been input in the system, which has generated a combination of brass instruments for each of the four timbral descriptors. The information about the instruments, playing techniques, and notes for each combination is detailed in Table 7.1.

The sequence of instrument combinations has been rendered as an audio file. Figure 7.6 shows a spectrogram of this output. Here, we can note a small amount of energy in the high frequencies for the first combination, whereas in the second and the fourth, there is a large presence of high frequencies. For the third instrument combination, the spectrogram displays high energy in several frequency bandwidths (represented by the red colours in horizontal lines) with short distances between them, a characteristic of the attribute *Roughness*.

Example 2 - From high brightness to low brightness For the second example, the sequence of timbral descriptors has been manually defined. Here, the system was asked to generate four string combinations for the attribute *Brightness*. Then, these combinations were organised from high brightness to low brightness. Table 7.2 details the instrument information for each combination and Figure 7.7 presents a spectrogram of the rendered output. We can see that the energy in the high frequencies decreases throughout the combinations of the sequence.

Example 3 - From dullness to roughness The third example presents a sequence of four combinations generated with brass instruments. The list of timbral descriptors consisted of two combinations for *Dullness* and two for *Roughness*. Then, the sequence was organised from a combination with a high dullness value to one with a lower value for the first half, then from a low roughness value combination to a higher roughness value combination for the second

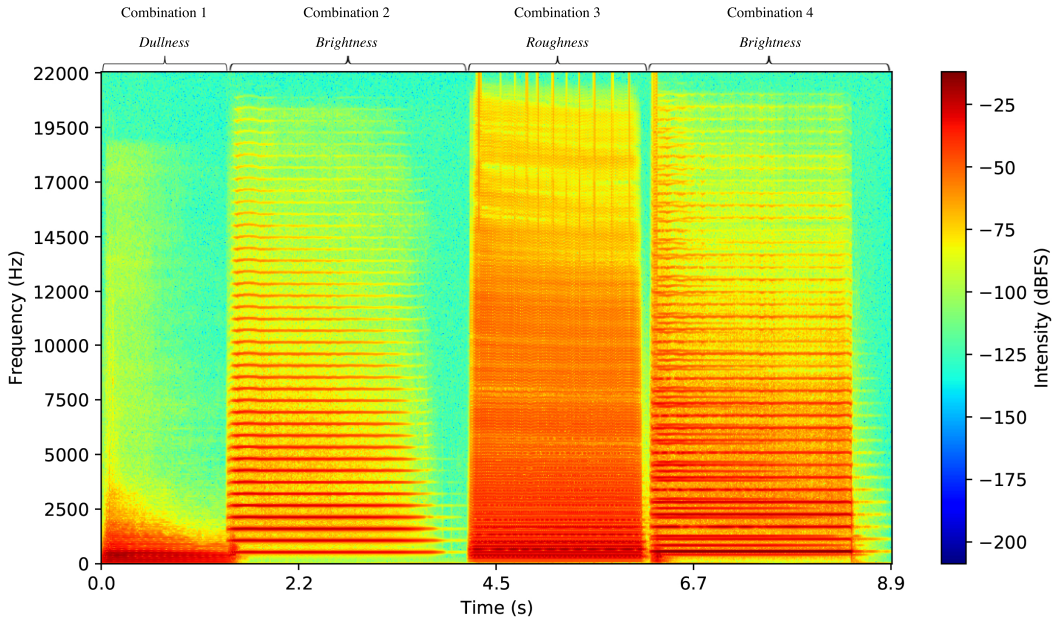

Figure 7.6: Spectrogram of the audio file generated for the sequence of instrument combination for *Example 1 - From audio file analysis*.

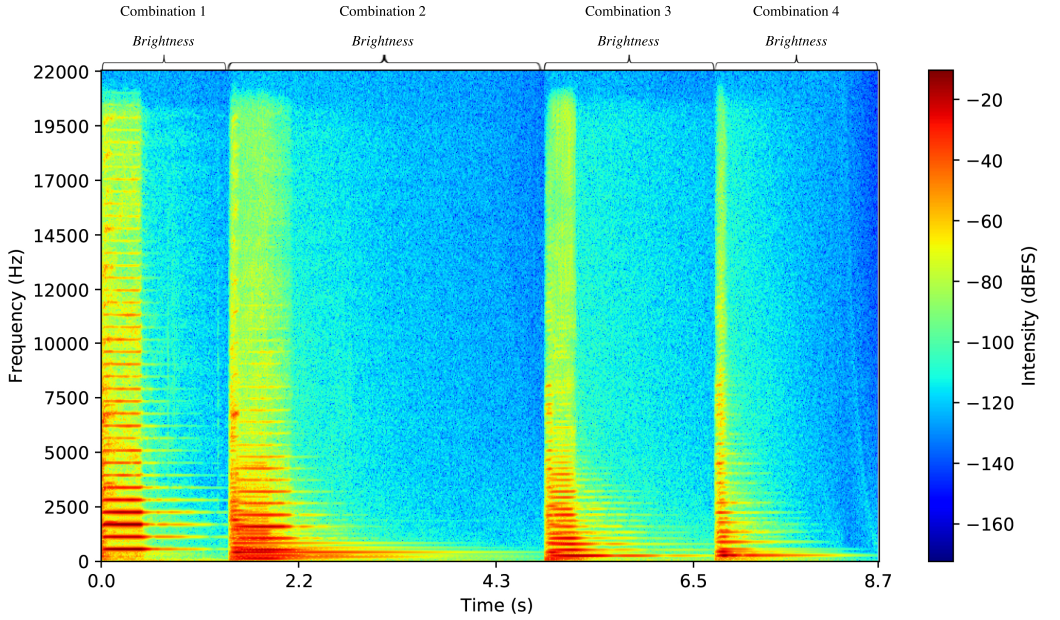

Figure 7.7: Spectrogram of the audio file generated for the sequence of instrument combinations for *Example 2 - From high brightness to low brightness*.

	Instrument	Playing technique	Note
Combination 1 - *Dullness*	Tuba	Decrescendo	G2
	Trombone	Staccato	B1
Combination 2 - *Brightness*	Trumpet	Sforzando	C5
	Horn	Staccato	E2
Combination 3 - *Roughness*	Trombone	Sforzando	E4
	Horn	Brassy	C3
Combination 4 - *Brightness*	Horn	Brassy	C♯5
	Trumpet	Sforzando	G♯4
	Trombone	Staccato	F3
	Tuba	Staccato	C2

Table 7.1: Details of the instrument combinations for the sequence of *Example 1 - From audio file analysis*.

	Instrument	Playing technique	Note
Combination 1 - *Brightness*	Violin	Staccato	A6
	Cello	Note-lasting	C♯5
Combination 2 - *Brightness*	Violin	Staccato	F6
	Viola	Note-lasting	C5
	Cello	Staccato	G♯3
	Bass	Note-lasting	C♯3
Combination 3 - *Brightness*	Violin	Staccato	G♯4
	Viola	Note-lasting	C4
	Bass	Pizzicato-secco	D♯3
Combination 4 - *Brightness*	Violin	Staccato	A4
	Bass	Staccato	C♯4

Table 7.2: Details of the instrument combinations for the sequence of *Example 2 - From high brightness to low brightness*.

part. The instrument information for each combination is detailed in Table 7.3. Similarly to the previous examples, the sequence have been rendered and a spectrogram of this audio file is shown in Figure 7.8. Here, we can note that the first combination have a low amount of high frequencies, represented by the lack of warm colours above 2500 Hz. The second combination presents further energy in the high frequencies than its predecessor, demonstrating a lesser dull sound. The other two combinations have a lot of energy in different frequency bandwidths,

	Instrument	Playing technique	Note
Combination 1 - *Dullness*	Tuba	Decrescendo	G2
	Trombone	Staccato	B1
Combination 2 - *Dullness*	Trumpet	Staccato	G♯4
	Tuba	Sforzando	E3
Combination 3 - *Roughness*	Trumpet	Sforzando	D♯4
	Horn	Brassy	G♯3
	Tuba	Sforzando	C3
Combination 4 - *Roughness*	Trumpet	Staccato	D♯5
	Horn	Brassy	E4
	Tuba	Sforzando	G♯3
	Trombone	Sforzando	B3

Table 7.3: Details of the instrument combinations for the sequence of *Example 3 - From dullness to roughness*.

with even higher energy in the bandwidths for the last combination. Furthermore, the distances between the bandwidths are short, corresponding to sounds having roughness qualities. This example illustrates the potential for creating sequences of instrument combinations going from one timbral attribute to another one.

The different examples presented in this section have demonstrated different approaches for generating sequences of instrument combinations. These sequencing methods can be the result of processing an audio file in order to estimate its timbral structure and then use this information as sets of verbal descriptor to generate. The other method utilises the same approach as discussed in the examples in Section 7.3, where descriptors are manually defined. Here, different types of ordering techniques can be set. For example, it is possible to generate combinations matching only one timbral quality and then order them in an ascending or descending order. Another approach is to use ordering methods to create transitions from one attribute to an other. The different modes utilised for the aforementioned examples introduce the basis for constructing sequences of instrument combinations.

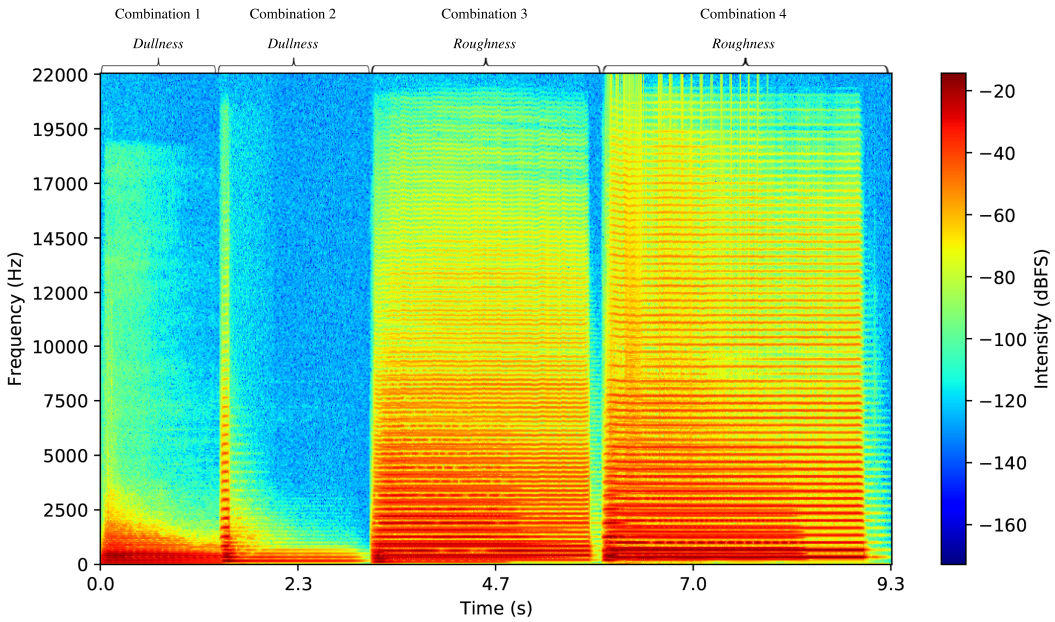

Figure 7.8: Spectrogram of the audio file generated for the sequence of instrument combinations for *Example 3 - From dullness to roughness*.

7.5 Chapter Conclusions

Several examples have been described in this chapter to illustrate the functioning and abilities of the trimbral driven generative instrument combinations system discussed in Chapter 6, which has incorporated the techniques identified from the research developments presented in Chapters 4 and 5. Such examples have also provided material to exemplify the methods for inputting combination's criteria into the generative system.

The first set of examples have discussed one generated combination for each of the five timbral descriptors. Here, different groups of string and brass instruments have been used for these combinations. Furthermore, the criteria about the number of instruments and descriptors were manually defined. The purpose of these examples was to illustrate the data handling and output techniques of the system. Moreover, the spectrograms of the rendered audio files depict the spectral contents of the different instrument combinations, highlighting their correlation with the implemented timbre estimation methods (Section 4.5).

After demonstrating the abilities for generating one instrument combination matching a specific timbral attribute, the second set of examples have discussed different methods for sequencing single events. One approach is to apply timbre estimations on audio files containing combinations of traditional Western instruments. This process is performed to retrieve the timbral structure and create a sequence of the estimated timbral descriptors, which will be subsequently utilised by the generative system. Another possible method is to generate several instrument combinations for one timbral descriptor and then sequence the outputs in ascending or descending order. This would offer a timbral evolution of the sonic qualities created by the instrument mixture. The last presented mode also utilises the ordering approach. However, the system is asked to generate combinations for two different timbral attributes. The first half of the sequence consists of combinations matching the first attribute in descending order. For the second half, the instrument combinations correspond to the second timbral descriptor and are sequenced in descending order. This approach proposes to create transitions from one attribute to an other.

This selection of examples has depicted a wide range of outputs highlighting their relationship to the input parameters. It has illustrated the type of instrument combinations that can be expected from the timbral driven generative system. As mentioned previously, the aim of this research is not to develop techniques to generate complete musical pieces. Instead, it is to provide methods and systems to aid composers by creating material to build on. It also aims to

establish techniques for the analysis and control of timbral qualities emerging from instrument combinations. The spectrograms have shown the correlation between the suggested acoustic calculations for each timbral descriptor (Section 4.5) and the spectral content of the instrument combinations output by the system. It highlights that the different methods for manipulating timbre properties from processing audio files have successfully been harnessed in the timbral driven generative system. However, with the nature of the sound created by combining instrument notes and the variations in an individual's audio perception, techniques merely based on signal processing are not enough to encompass all the properties related to the issue of instrumental timbre and timbral combinations. Suggested instrument combinations can present characteristics of a timbral attribute in its spectral content, but the correlation with its perception is not always compelling. Further investigations are required to address this issue and expand the identified techniques based on processing audio files towards standard methods for analysing and controlling perceptual properties of instrumental timbre combinations.

7.6 Chapter Summary

This chapter has presented different sets of instrument combinations generated using the system described in Section 6.3. These examples have been utilised to demonstrate and illustrate the abilities of the techniques and systems developed throughout the research project.

First, the text has detailed one example of instrument combination for each implemented timbral attribute. Here, the information about descriptors and instrument numbers and types were manually input. Spectrograms of the audio examples have been depicted in order to demonstrate the link between the spectal content of the instrument combinations and the implemented methods for timbre estimation. The second set of examples have focused on describing approaches for sequencing single events. Here, a mode involving timbral estimations of audio files to create sequences of descriptors has been described. The other modes propose to sequence instrument combinations matching one timbral quality in ascending or descending order and to create transitions between two attributes.

The examples detailed in this chapter have illustrated the research findings presented in Chapters 4, 5, and 6. They have also provided material to exemplify the methods for input information and the types of instrument combination outputs that can be produced by the timbral driven generative system.

8

Conclusions and Future Work

8.1 Chapter Overview

This chapter provides a summary of the different findings produced by the research developments presented within this thesis. Discussions on the different contributions are also provided, which give answers to the research questions listed in Section 1.3.1. The second part of this final chapter suggests avenues for future research. Such ideas could be further explored in order to expand the findings and applications of the present study. This section also suggests potential solutions to overcome some of the limitations that have been identified from the research developments discussed throughout this document.

The structure of this chapter is as follows:

- 8.2 - Research Conclusions
- 8.3 - Future Work

8.2 Research Conclusions

With the increase of accessibility and performances of the technology, composers, and more generally music enthusiasts, can experiment with a broad range of sophisticated tools to aid their compositional process. Such tools have successfully processed and manipulated many musical information, such as pitch, rhythm, and velocity. However, information related to timbre, especially in the context of combining instrument sounds, are yet to be completely harnessed in computing systems designed to aid compositional tasks. Some attempts at developing systems for composition practices involving several instruments have been carried out during the last decade, as discussed in Section 2.5. While some systems have proposed interesting approaches and methods, many challenges related to instrument combinations still require further investigations. Therefore, the aim of this research has been to establish techniques for the analysis and control of instrumental timbre and timbral combinations. This study has been conducted through the development of three different systems: a timbral ranking system, a timbral classification system, and a timbral driven generative instrument combinations system.

This investigation of harnessing timbre properties for use in computer-aided orchestration systems was underpinned by four research questions, which guided the evolution of the different implementations. The following sections provide a summary and conclusion for each of these, followed by a list of the different contributions to knowledge that were not directly related to the research questions.

8.2.1 RQ1: Which sonic element can be used to evaluate and represent the perceptual quality of the sound emerging from a mixture of instruments by processing audio recordings of instrument combinations?

Information to answer **RQ1** has been suggested by the discussions about human perception and timbre presented in Chapter 2. Here, it was identified that one attribute of sound is conveying perceptual information: timbre. This complex and multidimensional aspect of sound has been the subject of numerous studies for over a century, and continues to be an important area of research. Notwithstanding, Section 2.4 has highlighted that timbre can represent different properties. For this study, two main timbral paradigms have been examined. The first paradigm is related to instrument timbre, which are the properties that characterise an instrument's sound. This aspect is important in this study, due to processing audio samples that consist of recordings of instrument sounds. The second paradigm uses timbre to represent the

quality of a sound, which can be related to the notion of the 'colour' of a sound. Combining several instruments offers the possibilities to create unique textures. Here, timbre encompasses the different perceptual properties that allow the listener to experience these textures emerging from the instrumental mixtures, whether from polyphonic timbre, or from the phenomenon of timbre blending.

Discussions from Section 2.4 have highlighted that timbre characteristics can be retrieved by performing different acoustic calculations to measure various spectral and temporal information. Furthermore, due to timbre sometimes representing qualities of sounds, different words and adjectives, such as bright, dark, pure, or sharp, are often used to describe timbre characteristics. Such descriptors can be metaphors and analogies associated with the senses of vision and touch for example, representing perceptual qualities. Therefore, in order to alleviate the need for acoustic and psychoacoustic expertise, the use of verbal attributes to represent timbre properties were selected and not only their acoustic correlates. The objective was to make the presented approach accessible to a broad audience. However, even with the significant amount of studies on timbre, there are mixed conclusions in the definition of the acoustic cues correlating with the verbal attributes. In regards to polyphonic timbre, which is inherent to sounds created by combining instruments, the suggested methods are even more disparate, with some suggesting source separation prior to applying timbre estimations, while others stated that acoustic calculations could be applied to the combined sounds.

Due to the lack of agreed techniques for estimating timbre characteristics, the initial step of this research was to evaluate methods for calculating timbre properties from sources containing instrument combination sounds. Such audio sources involves performing acoustic calculations on polyphonic timbres. This investigation has been realised through the development of a timbral ranking system, designed to overcome the tedious task of going through the numerous solutions suggested by computer-aided orchestration systems (Section 2.5), which is one of their limitations. The objective of this initial system was to propose a ranking of the suggested solutions according to specific sound qualities, as defined in Section 4.3. Here, five verbal attributes, representing different sonic qualities, have been selected and used throughout this research: *breathiness*, *brightness*, *dullness*, *roughness*, and *warmth*. The rationale for selecting these verbal attributes has been discussed in Section 4.4. The methods to estimate the selected timbral attributes have been described in Section 4.5. Here, the calculations of the acoustic cues correlating with verbal attributes have been performed without applying a source separation process, but directly onto the instruments combinations as a single sound.

The system discussed in Chapter 4 has been designed to perform the acoustic analysis of each of the five selected verbal attributes directly onto audio files of recorded instrument sounds, in order to calculate timbral values. This process provided methods to compare the timbral values in order to define ranking of all the analysed audio files, as described in Section 4.6. A perceptual experiment, detailed in Section 4.7, has been conducted to evaluate the accuracy of the methods implemented for calculating the perceptual qualities of audio recordings of instrument combinations. The results of this perceptual experiment have suggested that the timbral values calculated from the estimations of specific acoustic cues were in correlation with the participants' ratings, thus following the human perception mechanism. Therefore, the development of this timbtral ranking system has contributed in establishing methods to evaluate specific perceptual qualities from sounds emerging from instrument mixtures by retrieving specific timbre characteristics. The results of the perceptual experiment have shown that the acoustic calculations, detailed in Section 4.5, could be applied directly onto global instrument combination sounds, without the need to perform a source separation process. This research development has also suggested that using verbal attributes to describe perceptual qualities could be used to represented timbre characteristics, which has been supported by the developments presented in Chapters 5 and 6.

8.2.1.1 RQ1 Final Remarks

A partial answer to **RQ1** had been suggested by the discussions about human perception and musical timbre, detailed in Chapter 2. However, **RQ1** has been addressed more specifically with the implementation of the timbral ranking system presented in Chapter 4. This system has proposed a solution to overcome the tedious task of having to listen to the numerous instrument combination solutions suggested by computer-aided orchestration systems, which was identified by experimenting with the *Orchids* program [14]. Furthermore, the use of timbre to estimate perceptual qualities from audio recordings of instrument sounds could be applied in the process of searching through audio file libraries. Estimating timbre characteristics could also be used in the decision rules of generative audio processes, as it has been harnessed in the developments discussed in Chapter 6.

8.2.2 RQ2: How to compare timbral values resulting from the analysis of different acoustic properties?

RQ2 is a follow up of the answer to **RQ1**, which has determined that timbre characteristics could be used to evaluate perceptual qualities from audio recordings of instrument sounds. Performing the different calculations of the acoustic properties resulted in obtaining timbral values for each verbal attribute. While it is possible to compare timbral values from the same verbal attribute, which was the approach selected in the timbral ranking system discussed in Chapter 4, it was not possible to compare the timbral values across the five verbal attributes, due to the different natures of acoustic properties. This limitation has been identified from the use of the timbral ranking system and the results of the perceptual experiment, reviewed in Section 4.7. Here, the system was still outputting a 'top' result, even if the audio files did not have the perceptual quality of an attribute, due to the algorithm comparing timbral values from a single verbal attribute, as mentioned in Section 5.2. Similarly to the definition of acoustic cues correlating with timbre characteristics, mixed conclusions have been suggested from the different studies trying to propose metrics for the calculated values, which means that no thresholds have been defined to determine the signification of the calculated timbral values.

Due to the lack of agreed metrics, it was necessary to establish a scale for each verbal attribute, in order to understand the characteristics of the calculated timbral values. The definition of the scales has been realised with the analysis of numerous audio files, consisting of recordings of orchestral pieces and combinations of instruments created by computing systems. The initial data gathering process consisted of an analysis of 50 orchestral pieces. Here, the audio files have been split into 1, 2, 3, 4, and 5 second audio samples, as described in Section 5.4.2. The rationale for splitting the long audio files is due to timbre properties involving temporal information, thus, analysing longer audio files would not provide meaningful values. This initial dataset has been further extended with the analysis of 205 extra orchestral pieces. This resulted with gathering over 236 000 values for each of the five verbal attributes, as detailed in Section 5.5.2.2. This large set of timbral values has contributed to establish a scale for each of the verbal attributes by performing statistical processes, which have also allowed for the rescaling of all the values, and therefore, have a comparative scale across the timbral values resulting from different acoustic calculations.

Section 5.4.4 has detailed the initial approach designed to define methods to classify instrument combinations sounds automatically according to their perceptual qualities. These meth-

ods were based on distance calculations and did not produce positive results. Here, distances measuring were not able to harness the timbral value characteristics for each attribute, and thus, such methods could not be used to compare calculated values in order to identify the dominant verbal attribute of an audio file. The results of this development suggested that individually comparing timbral values for each attribute was not providing complete information about the perceptual qualities' content of audio files, which may be due to the multidimensionality of timbre.

8.2.2.1 RQ2 Final Remarks

An answer to **RQ2** has been accomplished by performing a data gathering process in order to obtain significant data to understand the characteristics of the timbral values of each verbal attribute. Thanks to different statistical functions, it has been possible to determine a scale for each of the five selected attributes. With the establishment of scales, rescaling processes could be applied on the timbral values, which allowed the comparison of the data from different acoustic calculations, although the distance measuring methods utilised in Section 5.4 have not been able to harness the timbre characteristics from the calculated timbral values. The lack of agreed metrics makes the implementation and use of timbre properties in computing systems a challenging task, which is supported by the results of the distance calculation methods, for instance. Studies investigating this issue could benefit in establishing standards about the numbers calculated from estimating the different acoustic cues correlating with timbre properties.

8.2.3 RQ3: Which methods taken from the field of Artificial Intelligence (AI) could help in computationally analysing and identifying the perceptual properties of instrument combination sounds using values from musical timbre calculations?

RQ3 has been addressed with the developments presented in Chapter 5. Following the limitations of the timbral ranking system discussed in Chapter 4, which was suggesting a 'top' result despite a lack of the specific perceptual quality in the sound emerging from combining instrument notes, it was necessary to establish a method to identify the dominant quality of an audio file. The initial classification approach, mentioned in the answer to **RQ2**, and detailed in Section 5.4, has been based on distance calculations. However, the results of the different distance measures' testing, detailed in Section 5.4.3, have shown that this approach was not

capable of identifying the dominant perceptual quality from an instrument combination audio file using the calculated timbral values.

Following the negative performances of the initial classification approach, it was decided to investigate different machine learning algorithms in order to develop classification models based on timbral values. First, an unsupervised learning method was studied, designed to learn classification models by performing exploratory data analysis. The aim of this approach was to identify patterns or groupings in sets of unlabelled data. Therefore, a *k*-means algorithm was implemented, detailed in Section 5.5.2, and performed on the set of timbral values created to define the comparative scale proposing an answer to **RQ2**, and described in Section 5.5.2.2, which consisted of over 236 000 data samples. This machine learning method has been able to group similar timbral values into five clusters, representing the five verbal attributes. However, this method did not provide direct information on the names of the clusters, representing the different perceptual qualities, due to the use of unlabelled samples. Here, an evaluation of the clusters was required, which involved listening and labelling audio files in order to label the clusters. While the *k*-means algorithm has been able to suggest five clusters from set of timbral values, as discussed in Section 5.5.2.5, the extra cluster evaluation task and the lack of control in the learning process may be detrimental to the use of this method for classifying perceptual qualities of instrument combination sounds from timbral values.

From this observation, two supervised learning algorithms, where classification models are created from analysing sets of examples, have been investigated. Due to the algorithms requiring examples, the initial task was to create a training dataset. Here, 1250 samples were labelled, representing 250 examples for each of the five verbal attribute, as detailed in Section 5.5.3.2. The first supervised learning algorithm used a Support Vector Machines (SVM) method, described in Section 5.5.3.3, while the second was based on Artificial Neural Networks (ANNs), as detailed in Section 5.5.3.6. Both methods have scored successful classification rates (0.978 for SVM and 0.984 for ANNs), which has suggested that supervised learning algorithms were appropriate methods for automatically classifying sounds according to their perceptual qualities, using the timbral values calculated from specific acoustic properties. However, due to the learning mechanism of the supervised learning algorithm, a prior listening and labelling process is still required in order to create the training corpus.

The listening and labelling process required for the creation of the training dataset for the supervised learning algorithms was utilised to propose a solution to the challenges of harnessing aspects of human perception in computing systems. Here, this process has been integrated

in the development of a reinforced supervised learning algorithm, which has been detailed in Section 5.5.4. By putting a weight value on each training sample, the learning mechanisms emphasise on the training samples with the highest weight values, and therefore influence the learning process. This reinforced supervised learning method has been implemented with a SVM algorithm, which processes the weight values in its input data. A function has been designed to allow users to alter the weight values by creating their own training corpus and defining their own values, or by evaluating, and thus altering, the standard training dataset detailed in Section 5.5.3.2, which has been utilised for the supervised learning algorithms. This function provides a method to personalise the training models to the users' preferences, and therefore, create classification models harnessing individuals' audio perception.

The research developments presented in Chapter 5 have addressed **RQ3**, by suggesting that supervised learning algorithms based on SVM and ANNs methods have been able to generate classification models capable of identifying the perceptual quality of an instrument combination from its calculated timbral values. These developments contributed in defining an appropriate method for harnessing the values of specific timbre properties, offering a solution to the issue of comparing different types of values mentioned in the answer to **RQ2**. Furthermore, the development of a reinforced supervised learning method has proposed a solution to address the challenges of variations in human perception, by enabling the user to control aspects of the learning mechanisms. Such methods have offered an approach to overcome the need of listening to audio files in order to manually identify their perceptual qualities.

8.2.3.1 RQ3 Final Remarks

The development of the timbre classification system presented in Chapter 5 has provided information to address **RQ3**. This system was designed to overcome the limitation of the timbral ranking system discussed in Chapter 4, which was only individually comparing the timbral values from a defined set of audio files. The development of a method capable of automatically estimating the perceptual quality of instrument combination sounds could benefit different processes. For example, this method could offer a solution to overcome the time-consuming task of listening to a large audio sample database, by offering information about perceptual qualities and timbre content, which may not be provided by MIR tools, and thus, basing the search query on sonic qualities. Such developments could also be integrated into generative audio systems, enabling an automatic perceptual evaluation of the generated sounds, and therefore,

going towards adding a sonic perception aspect into AI methods designed for the generation of audio contents.

8.2.4 RQ4: How to incorporate timbre properties into algorithms designed to generate combinations of sampled instrument notes?

The implementation of the generative instrument combinations system presented in Chapter 6 has proposed an answer to **RQ4**. Building on the outcomes of the systems discussed in Chapters 4 and 5, developed to provide material to answer **RQ1**, **RQ2**, and **RQ3**, a computing system designed to suggest instrument combinations presenting specific timbral qualities has been implemented. This investigation has determined approaches to harness the method for calculating timbre characteristics from audio files and the techniques for automatic classification of instrument combination sounds according to their perceptual qualities.

The investigations into harnessing timbre properties have been conducted throughout the development of a generative system for combining string and brass instruments. The use of a smaller group of instruments has reduced the instrument combination space, but still involves the challenge of combining instrument timbres. The objective of this system was to generate combinations of notes that will be played simultaneously, which is an approach that encompasses the phenomenon of timbre blending and polyphonic timbre. The approach of this generative system was to distribute the notes across the different string and brass instruments, using verbal attributes describing perceptual qualities as the principal parameter for the generation of the note combination process.

The generative system used high-quality audio recordings of notes from four string instruments (bass, cello, viola, and violin) and four brass instruments (tuba, horn, trombone, and trumpet). Each instrument also had three different playing techniques: *note-lasting*, *pizzicato-secco*, and *staccato* for the string instruments, three techniques among *note-lasting*, *brassy*, *staccato*, *decrescendo*, and *sforzando* for each brass instrument. Within this sound database, detailed further in Section 6.3.2, the number of note combinations were already significant, suggesting a large search space. The use of a random generative process on this search space would result in notes combined arbitrarily, which may not necessarily be musically interesting, and involve large combinatorial possibilities. Therefore, in order to refine the search space, and propose combinations of notes following traditional Western musical elements, the rules for combining instrument notes were based on a chord framework, which involves combining notes to be played simultaneously. Here, three types of chords have been implemented: dyad,

triad, and seventh, including different variations for each of the three types, following the combinatorial rules detailed in Section 6.4.1. Generated chords were then rendered as audio files, in order to perform the timbre estimations, suggested in Chapter 4, to calculate the timbral values, which were then input into the timbre classification system, discussed in Chapter 5, and thus, being able to retrieve their perceptual qualities and to compare with the selected verbal attribute. While this method of audio generation and timbre estimations was able to evaluate the perceptual quality of each generated note combinations, the process was tedious and computationally expensive.

In order to address the audio generation and timbre estimation processes, two machine learning algorithms have been investigated to create regression models for predicting the timbral values directly from a set of information about the instruments, playing techniques, and notes that compose the generated chord. Here, a training dataset has been created by randomly generating numerous chord combinations for each of the three types, which have been generated as audio files in order to perform the timbre estimations function to calculate their timbral values. The performances of these machine learning methods, presented in Section 6.5, have suggested that the ANNs algorithm produced the best regression models. This technique provided a method for predicting timbral values directly from the instrument's information of the generated chords, allowing a quicker process for retrieving the chords' timbral values, and therefore accelerating the comparisons between the chords' perceptual qualities and the verbal attributes selected by the users.

The development of the generative system for string and brass instruments has contributed to harnessing the timbre estimations and timbre classification techniques suggested by the developments presented in Chapters 4 and 5 respectively. These techniques have been used to evaluate the perceptual qualities of chords generated by the system automatically. Here, the timbre properties, representing perceptual qualities, have been integrated in the decision rules of a search algorithm designed to generate combinations of sampled instrument notes, which are output only if they present specific timbral qualities. Moreover, the ANNs algorithm has contributed in creating regression models capable of predicting timbre characteristics directly from a set of instrument information, which addresses the large number of combinatorial possibilities. The implementation of this method for regression model has contributed in identifying the best parameters for the data mining note combination task.

8.2.4.1 RQ4 Final Remarks

The development of a computing system that generates combinations of instrument has offered an answer to **RQ4**, by using timbre properties as the fundamental criterion for a generative process designed to create combinations of notes to be played simultaneously by several instruments. This approach was based on processing audio samples and has successfully harnessed timbre properties by suggesting instrument combinations presenting acoustic characteristics of specific timbral attributes. However, further investigations are still required to enrich the signal processing techniques in order to completely encompass the perceptual properties of instrumental timbre combinations. Nevertheless, such methods could still be applied to other computer-aided composition systems, offering a control on some timbre phenomena that are inherent to the challenges of combining instrument sounds. Furthermore, the techniques developed for groups of string and brass instruments could then be applied in systems operating with larger ensembles, expanding the techniques for harnessing instrument timbre properties.

8.2.5 Other Contributions to Knowledge

The research presented within this thesis has been conducted to provide answers to the four research questions discussed previously. Nevertheless, this project has also produced contributions to knowledge in addition to the research questions listed in Section 1.3.1.

Research on timbre and its perception has been the subject of numerous perceptual experiments using various audio stimuli and methods (Section 2.4). However, the perceptual experiments conducted in this research project have utilised audio stimuli consisting of instrument combination sounds. Here, audio recordings of orchestral pieces and files generated by a computer-aided orchestration system have been used to represent instrument timbre combinations. Conducting perceptual experiments using this type of audio stimuli generated data about polyphonic timbre and timbre blending, which extends the understanding and processing of these two sonic phenomena.

Chapter 5 has discussed the research about developing timbre classification methods. Here, a data gathering process was implemented, which consisted of performing timbre analysis on 255 recordings of orchestral pieces. This resulted in calculating the timbral values of over 236,000 audio samples, and thus, establishing a database of timbral values from recordings of orchestral pieces. Another data gathering process has been performed for the creation of regression models, detailed in Chapter 6. Here, 60,000 chords were randomly generated, on

which the timbre calculations have been performed to retrieve the timbral values for each of the five timbral descriptors. Such datasets containing timbre values for polyphonic timbre and from instrument combination sounds could be utilised to further process and manipulate information about perceptual qualities represented by timbre properties.

Finally, while there is a significant amount of research using AI methods, as discussed in Chapter 3, the use of machine learning methods to process polyphonic timbre values has not been the focus of many works. The testing and parameter tuning process for creating classification models (Section 5.5) have provided information about the classification scores of different *k*-means, SVM, and ANNs functions. These processes have suggested parameters that have produced the best results for each of the three methods. Similar testing and parameter tuning techniques have been performed for the creation of models able to predict the timbral values of a chord directly from instrument information (Section 6.5). These processes provided data about the SVM and ANNs performances for generating regression models using datasets of instrument information and timbral values.

8.3 Future Work

This section discusses some areas for further investigation, which could extend the findings suggested by the different research developments presented within this thesis. Three main categories regroup different aspects to be investigated: notions related to timbre properties, investigations about instrument groups and approaches for their combinations, and interaction with other computing systems. The different avenues for future research are detailed in the following sections.

8.3.1 Timbre Properties

The study presented within this thesis has used timbre properties to represent perceptual qualities. Five verbal attributes (*breathiness*, *brightness*, *dullness*, *roughness*, and *warmth*) have been selected in the development of the different timbral systems discussed in Chapters 4, 5, and 6. These descriptors represent diverse sonic qualities, which can denote the type of sound desired by the users. However, these five verbal descriptors propose a limited range of sonic qualities. Therefore, one area of research could be to implement other attributes in the timbral systems presented in this thesis. This would offer the manipulation of a larger palette of perceptual qualities. Studies about semantic and timbre have proposed several verbal descriptors utilised to denote various qualities. A comprehensive list can be found in [117] for example, along with the corresponding acoustic cues for each attribute. The generative system discussed in Chapter 6 has harnessed the use of verbal attributes for guiding the generation of instruments notes to be played simultaneously, using combinatorial rules based on a chords framework. This type of note combination has been selected to represent the phenomena of timbre blending and polyphonic timbre. Research looking at harnessing timbre properties could benefit the types of solutions suggested by generative systems.

Another important area of research involves investigating timbre blending and polyphonic timbre resulting from the combination of several instruments. As mentioned in Chapters 2 and 4, studies investigating timbre have suggested mixed conclusions about the acoustic cues correlating with specific timbre characteristics, and their calculation techniques, which may be due to the variability in experiment methods and selected sound stimuli. This disparity is even larger when discussing polyphonic timbre, where some studies have suggested performing a source separation process and then apply the timbre estimations on each source, while other studies suggested that timbre calculation techniques could be applied directly to the global

sound. The perceptual experiment, reviewed in Section 4.7, has shown that the selected timbre estimation techniques used in Chapter 4, have been able to retrieve timbre characteristics by calculating the acoustic cues directly onto the instrument combination sounds. This lack of agreed techniques to estimate polyphonic timbre characteristics and their corresponding acoustic properties is a challenge for harnessing timbre properties in computing systems designed to aid compositional tasks involving several instruments. Furthermore, there is no agreed metrics for the values resulting from calculations of acoustic properties, as highlighted in Chapter 5. This lack of agreed metrics hinders the manipulation of values retrieved from timbre calculations. Therefore, research towards establishing standards for polyphonic timbre calculations and metrics, using standard experiment methods and similar sets of sound stimuli, would benefit the incorporation of timbre properties for addressing instrument combination challenges into computing systems.

8.3.2 Instruments Extension

The generative system discussed in Chapter 6 has utilised only groups of string and brass instruments to represent the challenge of composing for several instruments. However, the techniques developed to process these two types of instruments could be applied in systems operating with larger ensembles. In the presented generative system, this could be achieved by expanding the sound database with new audio recordings of instruments' notes, and by training the machine learning algorithms to create regression models based on the new set of instruments.

The second element that could be further developed is the note combination rules. Within the generative system, the rules have been based on a chord framework. These combinatorial rules, which are based on harmonic sets, have proposed combinations of notes that would sound more musical than applying a random process. However, a system generating only combinations of chords will propose a limited range of compositional material, which could be detrimental to the user's creativity. One area to investigate is to harness techniques to fuse instrumental sounds to create unique timbres that have been listed in treatises on orchestration, such as Piston's treatise [63] for example, where 'recipes' for combining instruments have been proposed. By incorporating these combinations, it would be possible to harness the available knowledge of instrumentation techniques, which could be used as foundations for generating combinations of notes. Furthermore, introducing rules about instrument constraints would provide controls for the horizontal generations. Here, the sequencing methods would

need to incorporate these rules to evaluate the playability between two generated instrument combinations. Therefore, expending the instrument combination methods, by adding different compositional frameworks and rules about instrumental technique constraints, would benefit greatly the musical possibilities of a computing system designed to aid compositional tasks.

8.3.3 Interaction with External Systems and Applications

Section 6.6 of this thesis has detailed the methods to define the sequences of instrument combinations to be created by the timbral driven generative system. Here, sets of timbral descriptors and number of instruments to combine can be input manually. They can also be the results of performing a timbre analysis of audio files in order to evaluate their timbral structure, which can then be utilised to sequence the generation's parameters.

A potential avenue for future work could be to focus on developing other methods to interact with the generative system. For instance, using *OSSIA Score*[1], an open-source intermedia sequencer [263], it is possible to construct scenarios on a time-line, which can represent the evolution of the instrument combination sequences. Figure 8.1 shows an example of a scenario, where the set of timbral descriptors is defined by the curves object at the top and the numbers of instruments to combine are defined by the curve at the bottom. Using their Python library [2], it is then possible to communicate with *OSSIA Score* and the timbral driven generative system to retrieve the information from the scenario. This option would allow the users to construct their sequences of combinations using a graphic interface. Furthermore, *OSSIA Score* integrates a conditional branching approach [264], which could be utilised to create interactive scenarios and use external MIDI hardware to control the generation's information input for example. Such developments would provide different approaches to create sequences of instrument combinations and offer a wider interaction between the system and the users.

[1] `https://ossia.io/`

[2] `https://github.com/OSSIA/libossia`

8. CONCLUSIONS AND FUTURE WORK

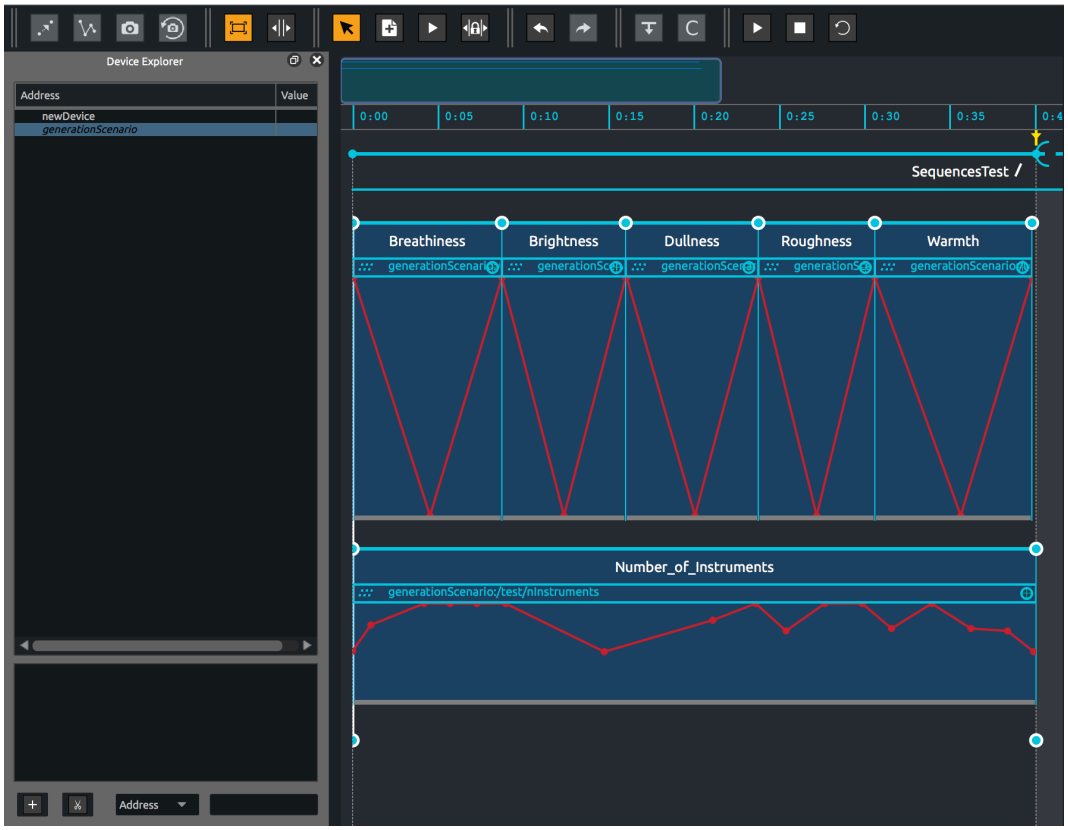

Figure 8.1: Screengrab of a scenario test in *OSSIA Score*.

Appendix A

Appendix A - Machine Learning Parameters Tuning Process for Classification Models

Appendix A displays the results of different classification functions' parameters tuning process mentioned in Section 5.5.3

A.1 Support Vector Machines (SVM) Parameters Tuning

	Precision	Recall	F1-score	Support
Breathiness	1.00	1.00	1.00	26
Brightness	1.00	1.00	1.00	24
Dullness	1.00	1.00	1.00	26
Roughness	0.95	0.92	0.94	24
Warmth	0.91	0.94	0.93	25
Avg / Total	0.97	0.97	0.97	125

Table A.1: Scores for the best `svm.SVC` function's parameters, with penalty parameter $C = 10$, kernel type $= rbf$, and rbf kernel coefficient $\gamma = 0.001$.

A. APPENDIX A - MACHINE LEARNING PARAMETERS TUNING PROCESS FOR CLASSIFICATION MODELS

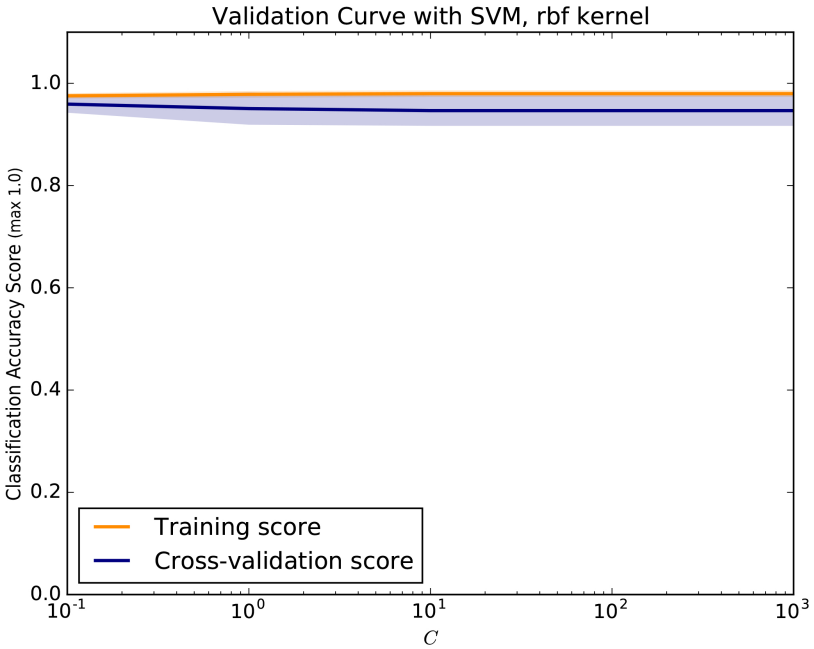

Figure A.1: Validation curve for the `svm.SVC` parameter C.

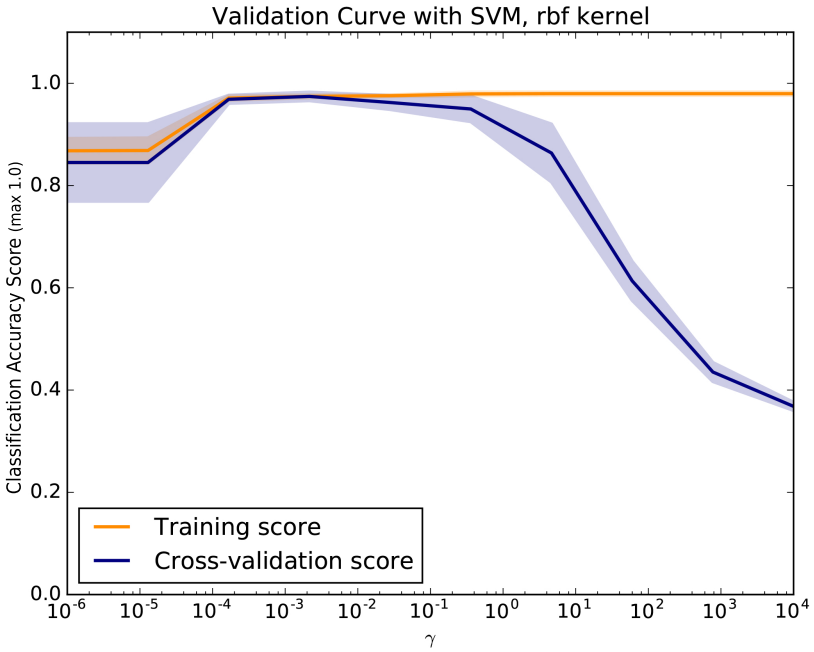

Figure A.2: Validation curve for the `svm.SVC` parameter γ.

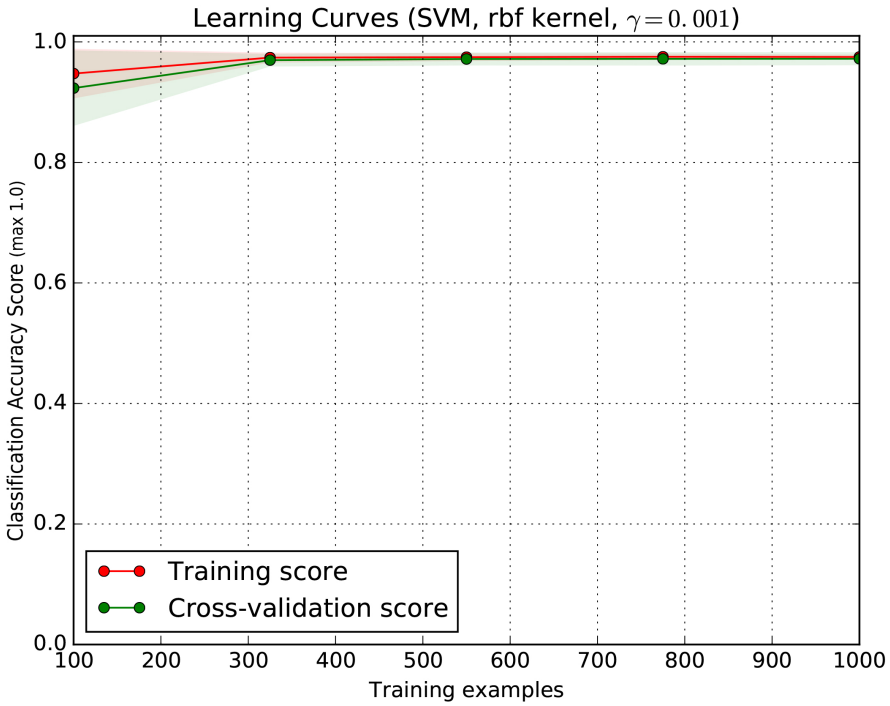

Figure A.3: Learning curves of the `svm.SVC` with parameters $kernel = rbf$, and $\gamma = 0.001$.

A.2 Artificial Neural Networks (ANNs) Parameters Tuning

	Precision	Recall	F1-score	Support
Breathiness	1.00	1.00	1.00	26
Brightness	1.00	1.00	1.00	24
Dullness	1.00	1.00	1.00	26
Roughness	0.96	0.92	0.94	24
Warmth	0.92	0.96	0.94	25
Avg / Total	0.98	0.98	0.98	125

Table A.2: Scores for the best `neural_network.MLPClassifier` function's parameters, with parameters $activation = identity$, $solver = adam$, and $learning\_rate = constant$

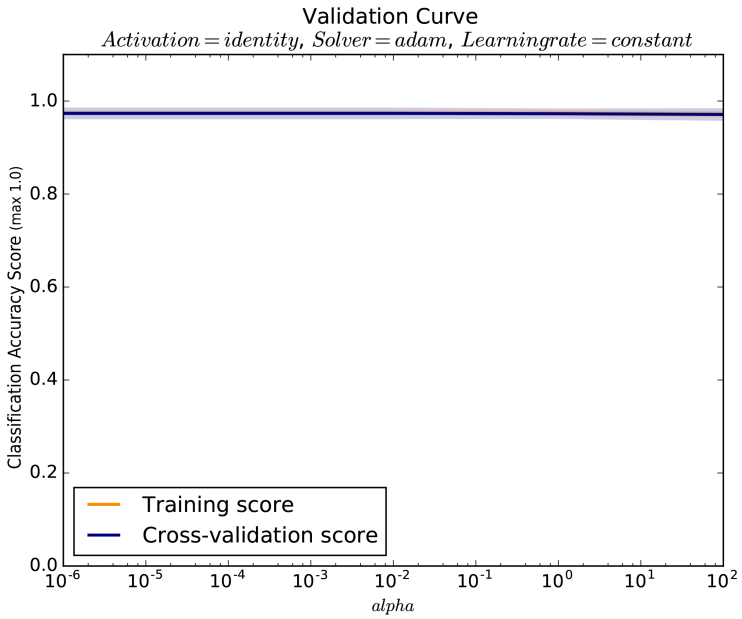

Figure A.4: Validation curve for the `neural_network.MLPClassifier` parameter $Alpha$.

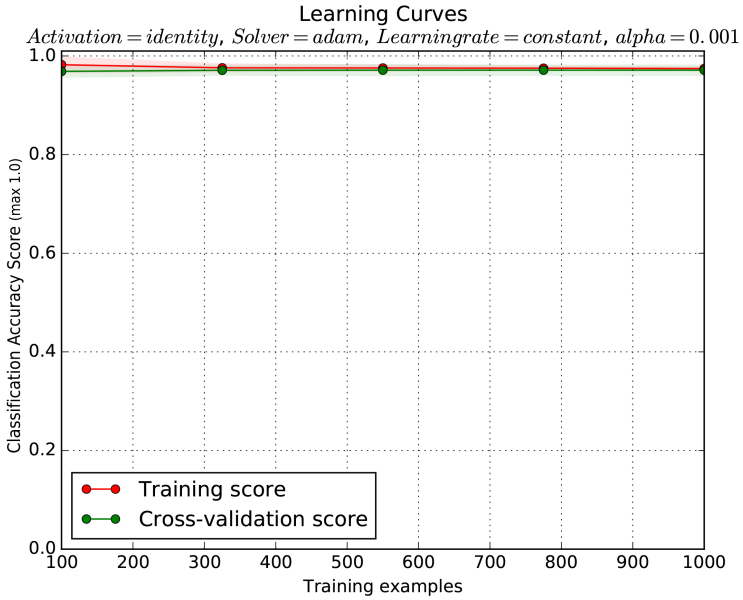

Figure A.5: Learning curves of the `neural_network.MLPClassifier` with parameters $activation = identity$, $solver = adam$, and $learning\ rate = constant$.

Reference List

[1] AMERICAN NATIONAL STANDARDS INSTITUTE. *Psychoacoustic terminology S3:20*. New York, NY: American National Standards Institute, 1973. [Cited on pages 2 and 23.]

[2] CAROL L KRUMHANSL. **Why is musical timbre so hard to understand**. *Structure and perception of electroacoustic sound and music*, **9**:43–53, 1989. [Cited on pages 2 and 23.]

[3] DENIS SMALLEY. **Defining timbre—refining timbre**. *Contemporary Music Review*, **10**(2):35–48, 1994. [Cited on page 2.]

[4] JOHN R PLATT AND RONALD J RACINE. **Effect of frequency, timbre, experience, and feedback on musical tuning skills**. *Perception & Psychophysics*, **38**(6):543–553, 1985. [Cited on pages 2 and 23.]

[5] ROBERT D MELARA AND LAWRENCE E MARKS. **Interaction among auditory dimensions: Timbre, pitch, and loudness**. *Perception & psychophysics*, **48**(2):169–178, 1990. [Cited on pages 2 and 23.]

[6] VALERIA C CARUSO AND EVAN BALABAN. **Pitch and timbre interfere when both are parametrically varied**. *PloS one*, **9**(1):e87065, 2014. [Cited on pages 2 and 23.]

[7] JAN FREDERIK SCHOUTEN. **The perception of timbre**. In *Reports of the 6th International Congress on Acoustics*, **76**, page 10, 1968. [Cited on pages 2 and 25.]

[8] JOHN M GREY. *An exploration of musical timbre*. PhD thesis, Stanford University, 1975. [Cited on page 2.]

[9] HUGO FASTL AND EBERHARD ZWICKER. *Psychoacoustics: Facts and models*, **22**. Springer Science and Business Media, 2007. [Cited on pages 2, 82, and 107.]

[10] HERMANN VON HELMHOLTZ. *On the sensations of tone as a physiological basis for the theory of music*. Dover Classics Of Science And Mathematics. Dover Publications, New York, 2nd English edition, 1954. [Cited on pages 2, 21, 22, 24, 26, and 28.]

[11] HUGUES DUFOURT. **Musique spectrale: pour une pratique des formes de l'énergie**. *Bicéphale*, (3):85–89, 1981. [Cited on pages 2, 34, and 41.]

[12] MING-HSUAN YANG, DAVID J KRIEGMAN, AND NARENDRA AHUJA. **Detecting faces in images: A survey**. *IEEE Transactions on pattern analysis and machine intelligence*, **24**(1):34–58, 2002. [Cited on pages 3 and 51.]

[13] FEI JIANG, YONG JIANG, HUI ZHI, YI DONG, HAO LI, SUFENG MA, YILONG WANG, QIANG DONG, HAIPENG SHEN, AND YONGJUN WANG. **Artificial intelligence in healthcare: past, present and future**. *Stroke and Vascular Neurology*, pages 1–14, 2017. [Cited on pages 3 and 51.]

[14] AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Towards intelligent orchestration systems**. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 671–681, Plymouth, UK, 2015. [Cited on pages 10, 39, and 204.]

[15] AURÉLIEN ANTOINE, DUNCAN WILLIAMS, AND EDUARDO R. MIRANDA. **Towards a timbre classification system for musical excerpts**. In *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP)*, London, UK, 2016. [Cited on page 10.]

[16] AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Musical acoustics, timbre, and computer–aided orchestration challenges**. In *Proceedings of the 2017 International Symposium on Musical Acoustics (ISMA)*, pages 151–154, Montreal, Canada, 2017. [Cited on page 10.]

[17] AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **A perceptually orientated approach for automatic classification of timbre content of orchestral excerpts**. *The Journal of the Acoustical Society of America*, **141**(5):3723, 2017. [Cited on page 10.]

[18] AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Computer generated orchestration: Towards using musical timbre in composition**. In *Pre-Proceedings of the 9th*

European Music Analysis Conference (EuroMAC 9 – IX CEAM), Strasbourg, France, 2017. [Cited on page 10.]

[19] EDUARDO R. MIRANDA, AURÉLIEN ANTOINE, JEAN-MICHAEL CELERIER, AND MYRIAM DESAINTE-CATHERINE. **i-Berlioz: interactive computer-aided orchestration with temporal control**. In *Proceedings of the 5th International Conference of New Musical Concepts (ICNMC)*, pages 45–60, Treviso, Italy, 2018. [Cited on page 11.]

[20] AURÉLIEN ANTOINE AND EDUARDO R. MIRANDA. **Predicting timbral and perceptual characteristics of orchestral instrument combinations**. *The Journal of the Acoustical Society of America*, **143**(3):1747, 2018. [Cited on page 11.]

[21] RICHARD C. ATKINSON, RICHARD J. HERRNSTEIN, GARDNER LINDZEY, AND R. DUNCAN LUCE, editors. *Stevens' handbook of experimental psychology*, **1**. John Wiley and Sons, New York, USA, 2nd edition, 1988. [Cited on pages 15 and 57.]

[22] GERAINT A WIGGINS. **Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music**. In *11th IEEE International Symposium on Multimedia (ISM09)*, pages 477–482. IEEE, 2009. [Cited on page 15.]

[23] FALLOPIO GABRIELE. *Observationes anatomicae*. Marcum Antonium Venise, Venice, 1561. [Cited on pages 16 and 42.]

[24] MARTIN MORTAZAVI, NIMER ADEEB, B LATIF, K WATANABE, DAS AMAN DEEP, CJ GRIESSENAUER, R. SHANE TUBBS, AND T FUKUSHIMA. **Gabriele Fallopio (1523–1562) and his contributions to the development of medicine and anatomy**. *Child's Nervous System*, pages 1–4, 2013. [Cited on pages 16 and 42.]

[25] DOMENICO COTUGNO. *De aquaeductibus auris humanae internae anatomica dissertatio*. Typographia Sanctae Tomae Aquinatis, Neopoli et Bononiae, 1761. [Cited on pages 16 and 42.]

[26] ERMANNO MANNI AND LAURA PETROSINI. **Domenico Cotugno, a pioneer in neurosciences**. *Journal of the History of the Neurosciences*, **6**(2):124–132, 1997. [Cited on pages 16 and 42.]

[27] ALFONSO CORTI. **Recherches sur l'organe de l'ouïe des Mammiferes**. *Zeitschrift für wissenschaftliche Zoologie*, **3**:109–169, 1851. [Cited on pages 16 and 42.]

[28] WALTER KLEY. **Alfonso Corti (1822–1876)–discoverer of the sensory end organ of hearing in Würzburg**. *ORL*, **48**(2):61–67, 1986. [Cited on pages 16 and 42.]

[29] GYÖRGY VON BÉKÉSY. *Experiments in hearing*. McGraw-Hill, 1960. (Translated by Ernest Glen Wever). [Cited on pages 16 and 26.]

[30] C. DANIEL GEISLER. *From sound to synapse: Physiology of the mammalian ear*. Oxford University Press, 1998. [Cited on page 16.]

[31] LARS CHITTKA AND AXEL BROCKMANN. **Perception space—the final frontier**. *PLoS Biol*, **3**(4):e137, 2005. [Cited on page 17.]

[32] JAMES PICKLES. **Auditory pathways: Anatomy and physiology**. In GASTONE CELESIA AND GREGORY HICKOK, editors, *H Handbook of clinical neurology: The human auditory system*, **129** of *3rd*, chapter 1, pages 2–25. Elsevier, 2015. [Cited on page 16.]

[33] NELSON YUAN-SHENG KIANG AND WILLIAM T PEAKE. **Physics and physiology of hearing**. In RICHARD C. ATKINSON, RICHARD J. HERRNSTEIN, GARDNER LINDZEY, AND R. DUNCAN LUCE, editors, *Stevens' handbook of experimental psychology*, **1**, pages 277–326. John Wiley and Sons, New York, 2nd edition, 1988. [Cited on page 16.]

[34] HENRY GRAY. *Anatomy of the human body*. Lea and Febiger, New York, 1918. [Cited on page 17.]

[35] IRA J HIRSH. **Auditory perception of temporal order**. *The Journal of the Acoustical Society of America*, **31**(6):759–767, 1959. [Cited on page 18.]

[36] IRA J. HIRSH. **Auditory perception and speech**. In RICHARD C. ATKINSON, RICHARD J. HERRNSTEIN, GARDNER LINDZEY, AND R. DUNCAN LUCE, editors, *Stevens' Handbook of Experimental Psychology*, **1**, pages 377–408. John Wiley and Sons, New York, 2nd edition, 1988. [Cited on page 18.]

[37] CHRISTOPHER J PLACK, ANDREW J OXENHAM, AND RICHARD R FAY. *Pitch: Neural coding and perception*, **24**. Springer Science and Business Media, 2006. [Cited on page 18.]

[38] HARVEY FLETCHER AND WILDEN A MUNSON. **Loudness, its definition, measurement and calculation**. *Bell Labs Technical Journal*, **12**(4):377–430, 1933. [Cited on pages 18 and 82.]

[39] STANLEY SMITH STEVENS. **A scale for the measurement of a psychological magnitude: loudness.** *Psychological Review*, **43**(5):405, 1936. [Cited on page 19.]

[40] STANLEY S STEVENS. **The measurement of loudness**. *The Journal of the Acoustical Society of America*, **27**(5):815–829, 1955. [Cited on page 19.]

[41] HERBERT WOODROW. **Time perception**. In STANLEY SMITH STEVENS, editor, *Handbook of Experimental Psychology*, pages 1224–1236. John Wiley and Sons, New York, 1951. [Cited on page 19.]

[42] PAUL FRAISSE. *The psychology of time*. Harper, New York, 1963. [Cited on page 19.]

[43] STUART ROSEN AND PETER HOWELL. *Signals and systems for speech and hearing*. Emerald Group Publishing Limited, 2nd edition, 2011. [Cited on page 19.]

[44] MAX V MATHEWS AND JOHN R PIERCE. **Harmony and nonharmonic partials**. *The Journal of the Acoustical Society of America*, **68**(5):1252–1257, 1980. [Cited on page 20.]

[45] BRIAN CJ MOORE, BRIAN R GLASBERG, AND ROBERT W PETERS. **Thresholds for hearing mistuned partials as separate tones in harmonic complexes**. *The Journal of the Acoustical Society of America*, **80**(2):479–483, 1986. [Cited on page 20.]

[46] RUDOLF A RASCH. **The perception of simultaneous notes such as in polyphonic music**. *Acta Acustica united with Acustica*, **40**(1):21–33, 1978. [Cited on page 20.]

[47] CJ DARWIN AND VALTER CIOCCA. **Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component**. *The Journal of the Acoustical Society of America*, **91**(6):3381–3390, 1992. [Cited on pages 20 and 31.]

[48] GOTTFRIED VON BISMARCK. **Timbre of steady sounds: A factorial investigation of its verbal attributes**. *Acta Acustica united with Acustica*, **30**(3):146–159, 1974. [Cited on pages 21 and 28.]

[49] WILLIAM FORDE THOMPSON AND LAURA-LEE BALKWILL. **Cross-cultural similarities and differences**. In PATRIK JUSLIN AND JOHN SLOBODA, editors, *Handbook of Music and Emotion: Theory, Research, Applications*, pages 755–788. Oxford University Press, New York, 2010. [Cited on page 21.]

[50] STEPHEN MCADAMS AND EMMANUEL BIGAND. *Thinking in sound: The cognitive psychology of human audition.* Oxford University Press, New York, 1993. [Cited on page 21.]

[51] MARI RIESS JONES, RICHARD R. FAY, AND ARTHUR N. POPPER, editors. *Music perception*, **36** of *Springer handbook of auditory research*. Springer Science and Business Media, 2010. [Cited on page 21.]

[52] DIANA DEUTSCH. *The psychology of music*. Academic Press, San Diego, US, 3rd edition, 2013. [Cited on page 21.]

[53] HECTOR BERLIOZ. *Grand traité d'instrumentation et d'orchestration modernes*. Henri Lemoine, Paris, 2e edition, 1855. [Cited on pages 22 and 23.]

[54] KENT KENNAN. *The technique of orchestration*. Prentice Hall, New Jersey, 1952. [Cited on page 22.]

[55] WAYNE SLAWSON. *Sound color*. Yank Gulch Music, 1985. [Cited on pages 22 and 26.]

[56] HARVEY FLETCHER. **Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure.** *Journal of the Acoustical Society of America*, 1934. [Cited on page 22.]

[57] CARL EMIL SEASHORE. *Psychology of music*. McGraw-Hill, New York, 1938. [Cited on page 22.]

[58] ALBERT S BREGMAN. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994. [Cited on page 23.]

[59] WILLIAM A SETHARES, ANDREW J MILNE, STEFAN TIEDJE, ANTHONY PRECHTL, AND JAMES PLAMONDON. **Spectral tools for dynamic tonality and audio morphing**. *Computer Music Journal*, **33**(2):71–84, 2009. [Cited on page 23.]

[60] ROBERT PRATT AND PHILIP E DOAK. **A subjective rating scale for timbre**. *Journal of Sound and Vibration*, **45**(3):317–328, 1976. [Cited on pages 23, 24, 28, 82, and 107.]

[61] HUGH MACDONALD ET AL. *Berlioz's orchestration treatise: a translation and commentary*. Cambridge University Press, 2002. [Cited on page 24.]

[62] NIKOLAY RIMSKY-KORSAKOV. *Principles of orchestration*. E. F. Kalmus, 1891. [Cited on page 24.]

[63] WALTER PISTON. *Orchestration*. WW Norton, 1955. [Cited on pages 24, 39, and 214.]

[64] ROGER A KENDALL AND EDWARD C CARTERETTE. **Verbal attributes of simultaneous wind instrument timbres: II. adjectives induced from Piston's orchestration**. *Music Perception: An Interdisciplinary Journal*, **10**(4):469–501, 1993. [Cited on page 24.]

[65] DAVID BUTLER. *The musician's guide to perception and cognition*. Schirmer Books, New York, 1992. [Cited on page 24.]

[66] REINIER PLOMP. **Timbre as a multidimensional attribute of complex tones**. In R. REINIER PLOMP AND G. F SMOORENBURG, editors, *Frequency analysis and periodicity detection in hearing*, pages 397–414. Sijthoff, 1970. [Cited on page 24.]

[67] STEPHEN HANDEL. **Timbre perception and auditory object identification**. In BRIAN C.J. MOORE, editor, *Hearing*, chapter 12, pages 425–461. Academic Press, San Diego, US, 1995. [Cited on pages 24 and 27.]

[68] STEPHEN HANDEL AND MOLLY L ERICKSON. **Sound source identification: The possible role of timbre transformations**. *Music Perception: An Interdisciplinary Journal*, **21**(4):587–610, 2004. [Cited on page 25.]

[69] AMERICAN STANDARDS ASSOCIATION (ASA). *Acoustical terminology, definition 12.9, timbre*, 1960. [Cited on page 25.]

[70] EDWIN GARRIGUES BORING. *Sensation and perception in the history of experimental psychology*. Appleton-Century-Crofts, 1942. [Cited on page 26.]

[71] RC MATHES AND RL MILLER. **Phase effects in monaural perception**. *The Journal of the Acoustical Society of America*, **19**(5):780–797, 1947. [Cited on page 26.]

[72] REINIER PLOMP AND HERMAN J. M. STEENEKEN, STEENEKEN. **Effect of phase on the timbre of complex tones**. *The Journal of the Acoustical Society of America*, **46**(2B):409–421, 1969. [Cited on page 26.]

[73] JOHN M GREY AND JOHN W GORDON. **Perceptual effects of spectral modifications on musical timbres**. *The Journal of the Acoustical Society of America*, **63**(5):1493–1500, 1978. [Cited on pages 26 and 27.]

[74] STEPHEN MCADAMS. **Perspectives on the contribution of timbre to musical structure**. *Computer Music Journal*, **23**(3):85–102, 1999. [Cited on pages 26 and 27.]

[75] STEPHEN LAKATOS. **A common perceptual space for harmonic and percussive timbres**. *Perception and psychophysics*, **62**(7):1426–1439, 2000. [Cited on pages 26, 28, and 29.]

[76] JEAN-CLAUDE RISSET AND DAVID L WESSEL. **Exploration of timbre by analysis and synthesis**. In DIANA DEUTSCH, editor, *The psychology of music*, pages 113–169. Academic Press, San Diego, US, 2nd edition, 1999. [Cited on page 26.]

[77] WOLFGANG KÖHLER. **Akustische Untersuchungen. II**. *Zeitschrift für Psychologie*, **58**:59–140, 1910. [Cited on page 26.]

[78] CARL STUMPF. *Die sprachlaute*. Springer Verlag, Berlin and New York, 1926. [Cited on pages 26 and 27.]

[79] JAMES HOPWOOD JEANS. *Science and music*. Courier Corporation, 1937. [Cited on page 26.]

[80] A WAYNE SLAWSON. **Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency**. *The Journal of the Acoustical Society of America*, **43**(1):87–101, 1968. [Cited on page 26.]

[81] REINIER PLOMP AND HERMAN J. M. STEENEKEN. **Pitch versus timbre**. In *Seventh International Congress on Acoustics, Budapest*, pages 378–380, 1971. [Cited on page 26.]

[82] CAROL L KRUMHANSL AND PAUL IVERSON. **Perceptual interactions between musical pitch and timbre.** *Journal of Experimental Psychology: Human Perception and Performance*, **18**(3):739, 1992. [Cited on page 27.]

[83] EL SALDANHA AND JOHN F CORSO. **Timbre cues and the identification of musical instruments**. *The Journal of the Acoustical Society of America*, **36**(11):2021–2026, 1964. [Cited on page 27.]

[84] WH GEORGE. **A sound reversal technique applied to the study of tone quality**. *Acta Acustica united with Acustica*, **4**(1):224–225, 1954. [Cited on page 27.]

[85] KENNETH W BERGER. **Some factors in the recognition of timbre**. *The Journal of the Acoustical Society of America*, **36**(10):1888–1891, 1964. [Cited on page 27.]

[86] REINIER PLOMP. *Aspects of tone sensation: A psychophysical study*. Academic Press, 1976. [Cited on page 27.]

[87] JOHN M GREY. **Multidimensional perceptual scaling of musical timbres**. *The Journal of the Acoustical Society of America*, **61**(5):1270–1277, 1977. [Cited on pages 27 and 28.]

[88] DAVID L WESSEL. **Timbre space as a musical control structure**. *Computer music journal*, pages 45–52, 1979. [Cited on page 27.]

[89] ROGER A KENDALL AND EDWARD C CARTERETTE. **Perceptual scaling of simultaneous wind instrument timbres**. *Music Perception: An Interdisciplinary Journal*, **8**(4):369–404, 1991. [Cited on pages 27, 30, and 31.]

[90] PAUL IVERSON AND CAROL L KRUMHANSL. **Isolating the dynamic attributes of musical timbrea**. *The Journal of the Acoustical Society of America*, **94**(5):2595–2603, 1993. [Cited on pages 27 and 28.]

[91] JOCHEN KRIMPHOFF, STEPHEN MCADAMS, AND SUZANNE WINSBERG. **Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique**. *Le Journal de Physique IV*, **4**(C5):C5–625, 1994. [Cited on pages 27 and 28.]

[92] STEPHEN MCADAMS, SUZANNE WINSBERG, SOPHIE DONNADIEU, GEERT SOETE, AND JOCHEN KRIMPHOFF. **Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes**. *Psychological research*, **58**(3):177–192, 1995. [Cited on pages 27, 28, 29, and 31.]

[93] ROGER N SHEPARD. **The analysis of proximities: Multidimensional scaling with an unknown distance function. I.** *Psychometrika*, **27**(2):125–140, 1962. [Cited on page 27.]

[94] JOSEPH B KRUSKAL. **Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis**. *Psychometrika*, **29**(1):1–27, 1964. [Cited on page 27.]

[95] INGWER BORG AND PATRICK JF GROENEN. *Modern multidimensional scaling: Theory and applications*. Springer Science and Business Media, 2005. [Cited on page 27.]

[96] STEPHEN MCADAMS. **Recognition of auditory sound sources and events**. In STEPHEN MCADAMS AND EMMANUEL BIGAND, editors, *Thinking in sound: the cognitive psychology of human audition*, chapter 6, pages 146–198. Oxford University Press, New York, 1993. [Cited on page 27.]

[97] JOHN M HAJDA, ROGER A KENDALL, EDWARD C CARTERETTE, AND MICHAEL L HARSHBERGER. **Methodological issues in timbre research.** In IRÈNE DELIÈGE AND JOHN SLOBODA, editors, *Perception and cognition of music*, chapter 12, pages 253—306. Psychology Press, Hove, UK, 1997. [Cited on pages 27, 28, and 32.]

[98] STEPHEN MCADAMS. **Musical timbre perception**. In DIANA DEUTSCH, editor, *The psychology of music*, pages 35–67. Academic Press, San Diego, US, 3rd edition, 2013. [Cited on page 27.]

[99] TAFFETA M ELLIOTT, LIBERTY S HAMILTON, AND FRÉDÉRIC E THEUNISSEN. **Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones**. *The Journal of the Acoustical Society of America*, **133**(1):389–404, 2013. [Cited on pages 27 and 29.]

[100] EMERY SCHUBERT, JOE WOLFE, AND ALEX TARNOPOLSKY. **Spectral centroid and timbre in complex, multiple instrumental textures**. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*, pages 112–116, 2004. [Cited on pages 28, 77, 81, and 106.]

[101] ANNE CACLIN, STEPHEN MCADAMS, BENNETT K SMITH, AND SUZANNE WINSBERG. **Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones**. *The Journal of the Acoustical Society of America*, **118**(1):471–482, 2005. [Cited on page 28.]

[102] EMERY SCHUBERT AND JOE WOLFE. **Does timbral brightness scale with frequency and spectral centroid?** *Acta acustica united with acustica*, **92**(5):820–825, 2006. [Cited on page 28.]

[103] WILLIAM HEIL LICHTE. **Attributes of complex tones.** *Journal of Experimental Psychology*, **28**(6):455, 1941. [Cited on page 28.]

[104] GOTTFRIED VON BISMARCK. **Sharpness as an attribute of the timbre of steady sounds**. *Acta Acustica united with Acustica*, **30**(3):159–172, 1974. [Cited on pages 28 and 79.]

[105] ERNST TERHARDT. **On the perception of periodic sound fluctuations (roughness)**. *Acta Acustica united with Acustica*, **30**(4):201–213, 1974. [Cited on pages 28, 79, 82, and 107.]

[106] ARNE NYKÄNEN AND ÖRJAN JOHANSSON. **Development of a language for specifying saxophone timbre**. In *Stockholm Music Acoustics Conference*, pages 647–650, Stockholm, Sweden, 2003. [Cited on page 28.]

[107] GRAHAM DARKE. **Assessment of timbre using verbal attributes**. In CAROLINE TRAUBE AND SERGE LACASSE, editors, *Second Conference on Interdisciplinary Musicology*, pages 1–12, Montreal, Canada, 2005. [Cited on pages 28 and 79.]

[108] ALASTAIR C DISLEY, DAVID M HOWARD, AND ANDY D HUNT. **Timbral description of musical instruments**. In *International Conference on Music Perception and Cognition*, pages 61–68, Bologna, Italy, 2006. [Cited on pages 28, 77, 81, and 106.]

[109] KAI SIEDENBURG, ICHIRO FUJINAGA, AND STEPHEN MCADAMS. **A comparison of approaches to timbre descriptors in music information retrieval and music psychology**. *Journal of New Music Research*, **45**(1):27–41, 2016. [Cited on page 28.]

[110] ROGER A KENDALL AND EDWARD C CARTERETTE. **Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives**. *Music Perception*, pages 445–467, 1993. [Cited on page 28.]

[111] CHARLES E OSGOOD, GEORGE J SUCI, AND PERCY H TANNENBAUM. *The measurement of meaning*. University of Illinois Press, Urbana, USA, 1957. [Cited on page 28.]

[112] ALEX GOUNAROPOULOS AND COLIN JOHNSON. **Synthesising timbres and timbre-changes from adjectives/adverbs**. In *EvoWorkshops 2006: Applications of Evolutionary Computing*, pages 664–675, Budapest, Hungary, 2006. Springer. [Cited on page 28.]

[113] EWA LUKASIK. **Towards timbre-driven semantic retrieval of violins**. In *Fifth International Conference on Intelligent Systems Design and Applications*, pages 55–60, Wroclaw, Poland, 2005. IEEE. [Cited on page 28.]

[114] WILLIAM A SETHARES. *Tuning, timbre, spectrum, scale*. Springer, 1998. [Cited on pages 28 and 86.]

[115] JAMES W BEAUCHAMP. **Synthesis by spectral amplitude and "Brightness" matching of analyzed musical instrument tones**. *Journal of the Audio Engineering Society*, **30**(6):396–406, 1982. [Cited on page 28.]

[116] JENS HJORTKJÆR AND STEPHEN MCADAMS. **Spectral and temporal cues for perception of material and action categories in impacted sound sources**. *The Journal of the Acoustical Society of America*, **140**(1):409–420, 2016. [Cited on page 28.]

[117] DUNCAN WILLIAMS. *Towards a timbre morpher*. PhD thesis, University of Surrey, Surrey, UK, 2010. [Cited on pages 28, 79, and 213.]

[118] ERICH SCHUMANN. *Die Physik der Klangfarben*. PhD thesis, Humboldt University, Berlin, Germany, 1929. [Cited on page 29.]

[119] CHRISTOPH REUTER. **Stream segregation and formant areas**. In *European Society for the Cognitive Sciences of Music Conference (ESCOM)*, 2003. [Cited on page 29.]

[120] ROGER A KENDALL. **The role of acoustic signal partitions in listener categorization of musical phrases**. *Music Perception: An Interdisciplinary Journal*, **4**(2):185–213, 1986. [Cited on pages 29 and 32.]

[121] GREGORY J SANDELL AND MICHAEL CHRONOPOULOS. **Identifying musical instruments from multiple versus single notes**. *The Journal of the Acoustical Society of America*, **100**:2752, 1996. [Cited on page 29.]

[122] KEITH DANA MARTIN. *Sound-source recognition: A theory and computational model*. PhD thesis, Massachusetts Institute of Technology (MIT), Boston, USA, 1999. [Cited on page 29.]

[123] ASHA SRINIVASAN, DAVID SULLIVAN, AND ICHIRO FUJINAGA. **Recognition of isolated instrument tones by conservatory students**. In *Proceedings of the International Conference on Music Perception and Cognition*, pages 17–21, 2002. [Cited on page 29.]

[124] JUDITH C BROWN. **Computer identification of musical instruments using pattern recognition with cepstral coefficients as features**. *The Journal of the Acoustical Society of America*, **105**(3):1933–1941, 1999. [Cited on page 29.]

[125] KEITH D MARTIN. **Toward automatic sound source recognition: identifying musical instruments**. *NATO computational hearing advanced study institute*, 1998. [Cited on page 29.]

[126] ANTTI ERONEN AND ANSSI KLAPURI. **Musical instrument recognition using cepstral coefficients and temporal features**. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **2**, pages 753–756. IEEE, 2000. [Cited on page 29.]

[127] PERFECTO HERRERA-BOYER, GEOFFROY PEETERS, AND SHLOMO DUBNOV. **Automatic classification of musical instrument sounds**. *Journal of New Music Research*, **32**(1):3–21, 2003. [Cited on page 29.]

[128] STEVEN DAVIS AND PAUL MERMELSTEIN. **Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences**. *IEEE transactions on acoustics, speech, and signal processing*, **28**(4):357–366, 1980. [Cited on page 29.]

[129] GAËL RICHARD, SHIVA SUNDARAM, AND SHRIKANTH NARAYANAN. **An overview on perceptually motivated audio indexing and classification**. *Proceedings of the IEEE*, **101**(9):1939–1954, 2013. [Cited on page 29.]

[130] ANTTI ERONEN. **Comparison of features for musical instrument recognition**. In *Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 19–22. IEEE, 2001. [Cited on page 29.]

[131] KAILASH PATIL, DANIEL PRESSNITZER, SHIHAB SHAMMA, AND MOUNYA ELHILALI. **Music in our ears: The biological bases of musical timbre perception**. *PLoS Comput Biol*, **8**(11):e1002759, 2012. [Cited on page 29.]

[132] KAILASH PATIL AND MOUNYA ELHILALI. **Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases**. *EURASIP Journal on Audio, Speech, and Music Processing*, **2015**(1):27, 2015. [Cited on page 29.]

[133] EDGAR HEMERY AND JEAN-JULIEN AUCOUTURIER. **One hundred ways to process time, frequency, rate and scale in the central auditory system: A pattern-recognition meta-analysis**. *Frontiers in computational neuroscience*, **9**, 2015. [Cited on page 29.]

[134] TAFFETA M ELLIOTT AND FRÉDÉRIC E THEUNISSEN. **The modulation transfer function for speech intelligibility**. *PLoS comput biol*, **5**(3):e1000302, 2009. [Cited on page 29.]

[135] ETIENNE THORET, PHILIPPE DEPALLE, AND STEPHEN MCADAMS. **Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments**. *The Journal of the Acoustical Society of America*, **140**(6):478–483, 2016. [Cited on page 29.]

[136] ETIENNE THORET, PHILIPPE DEPALLE, AND STEPHEN MCADAMS. **Perceptually salient regions of the modulation power spectrum for musical instrument identification**. *Frontiers in Psychology*, **8**, 2017. [Cited on page 29.]

[137] FERDINAND FUHRMANN. *Automatic musical instrument recognition from polyphonic music audio signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2012. [Cited on pages 29 and 32.]

[138] JEAN-JULIEN AUCOUTURIER, FRANÇOIS PACHET, AND MARK SANDLER. **"The way it sounds": timbre models for analysis and retrieval of music signals**. *IEEE Transactions on Multimedia*, **7**(6):1028–1035, 2005. [Cited on page 30.]

[139] JEAN-JULIEN AUCOUTURIER. *Dix expériences sur la modélisation du timbre polyphonique*. PhD thesis, Université Paris 6, Paris, France, 2006. [Cited on pages 30, 31, and 32.]

[140] ROBERT O GJERDINGEN AND DAVID PERROTT. **Scanning the dial: The rapid recognition of music genres**. *Journal of New Music Research*, **37**(2):93–100, 2008. [Cited on page 30.]

[141] MARK D PLUMBLEY, SAMER A ABDALLAH, JUAN PABLO BELLO, MIKE E DAVIES, GIULIANO MONTI, AND MARK B SANDLER. **Automatic music transcription and audio source separation**. *Cybernetics and Systems*, **33**(6):603–627, 2002. [Cited on page 30.]

[142] GIANPAOLO EVANGELISTA, SYLVAIN MARCHAND, MARK PLUMBLEY, AND EMMANUEL VINCENT. **Sound source separation**. In UDO ZÖLZER, editor, *DAFX-Digital Audio Effects*, chapter 14, pages 551–558. John Wiley and Sons, 2nd edition, 2011. [Cited on page 30.]

[143] GREGORY JOHN SANDELL. *Concurrent timbres in orchestration: A perceptual study of factors determining 'blend'*. PhD thesis, Northwestern University, 1991. [Cited on page 31.]

[144] GREGORY JOHN SANDELL. **Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration**. *Music Perception: An Interdisciplinary Journal*, **13**(2):209–246, 1995. [Cited on page 31.]

[145] DAMIEN TARDIEU AND STEPHEN MCADAMS. **Perception of dyads of impulsive and sustained instrument sounds**. *Music Perception: An Interdisciplinary Journal*, **30**(2):117–128, 2012. [Cited on page 31.]

[146] PAUL IVERSON. **Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes.** *Journal of Experimental Psychology: Human Perception and Performance*, **21**(4):751, 1995. [Cited on page 31.]

[147] VINOO ALLURI AND PETRI TOIVIAINEN. **Exploring perceptual and acoustical correlates of polyphonic timbre**. *Music Perception: An Interdisciplinary Journal*, **27**(3):223–242, 2010. [Cited on page 32.]

[148] VINOO ALLURI. *Acoustic, neural, and perceptual correlates of polyphonic timbre*. PhD thesis, University of Jyväskylä, Jyväskylä, Finland, 2012. [Cited on page 32.]

[149] GÉRARD ASSAYAG, CAMILO RUEDA, MIKAEL LAURSON, CARLOS AGON, AND OLIVIER DELERUE. **Computer-assisted composition at IRCAM: From PatchWork to OpenMusic**. *Computer Music Journal*, **23**(3):59–72, 1999. [Cited on page 34.]

[150] CLARENCE BARLOW. **On the spectral analysis of speech for subsequent resynthesis by acoustic instruments**. In *Forum phoneticum*, **66**, pages 183–190. Hector, 1998. [Cited on pages 34 and 41.]

[151] TOM ROJO POLLER. **Clarence Barlow's: Technique of 'synthrumentation' and its use in im januar am nil**. *Tempo*, **69**(271):7–23, January 2015. [Cited on page 34.]

[152] CLAUDY MALHERBE. **Locus: rien n'aura eu lieu que le lieu**. In *The OM Composer's Book*, **2**. Editions Delatour, France, 2008. [Cited on page 34.]

[153] CARLOS AGON, GÉRARD ASSAYAG, JOSHUA FINEBERG, AND CAMILO RUEDA. **Kant: A critique of pure quantification**. In *Proceedings of the International Computer Music Conference*, pages 52–9, 1994. [Cited on page 35.]

[154] YAN MARESZ. **On computer-assisted orchestration**. *Contemporary Music Review*, **32**(1):99–109, 2013. [Cited on pages 35, 36, 37, and 41.]

[155] GEORG HAJDU. **Macaque – A tool for spectral processing and transcription**. In *Proceedings of the International Conference on Technologies for Music Notation and Representation (Tenor)*, 2017. [Cited on page 35.]

[156] FRANÇOIS ROSE AND JAMES HETRICK. **Spectral analysis as a ressource for contemporary orchestration technique**. In *Proceedings of Conference on Interdisciplinary Musicology*, **2005**, 2005. [Cited on pages 35 and 41.]

[157] DAVID PSENICKA. **Sporch: An algorithm for orchestration based on spectral analyses of recorded sounds**. In *Proceedings of International Computer Music Conference (ICMC)*, page 184, 2003. [Cited on pages 36 and 41.]

[158] THOMAS A HUMMEL. **Simulation of human voice timbre by orchestration of acoustic music instruments**. In *Proceedings of International Computer Music Conference (ICMC)*, page 185, 2005. [Cited on pages 36 and 41.]

[159] DAMIEN TARDIEU. *Modèles d'instruments pour l'aide à l'orchestration*. PhD thesis, Université Pierre et Marie Curie (UPMC) et IRCAM, Paris, 2008. [Cited on pages 36 and 41.]

[160] GRÉGOIRE CARPENTIER. *Approche computationnelle de l'orchestration musicale*. PhD thesis, Université Pierre et Marie Curie (UPMC) et IRCAM, Paris, 2008. [Cited on pages 36 and 41.]

[161] DAMIEN TARDIEU, GRÉGOIRE CARPENTIER, AND XAVIER RODET. **Computer-aided orchestration based on probabilistic instruments models and genetic exploration**. In *Proceedings of International Computer Music Conference, Copenhagen, Denmark*, 2007. [Cited on page 37.]

[162] GRÉGOIRE CARPENTIER AND JEAN BRESSON. **Interacting with symbol, sound, and feature spaces in orchidée, a computer-aided orchestration environment**. *Computer Music Journal*, **34**(1):10–27, 2010. [Cited on page 37.]

[163] GILBERT NOUNO, ARSHIA CONT, GRÉGOIRE CARPENTIER, AND JONATHAN HARVEY. **Making an orchestra speak**. In *Proceedings of the Sound and Music Computing Conference (SMC)*, Porto, Portugal, 2009. [Cited on page 37.]

[164] PHILIPPE ESLING. *Multiobjective time series matching and classification*. PhD thesis, Université Pierre et Marie Curie (UPMC) et IRCAM, Paris, 2012. [Cited on pages 37, 39, and 41.]

[165] GEOFFROY PEETERS. **A large set of audio features for sound description (similarity and classification) in the CUIDADO project**. Technical report, CUIDADO I.S.T., Ircam, 2004. [Cited on page 37.]

[166] PHILIPPE ESLING AND ANTOINE BOUCHEREAU. *Orchids: Abstract and temporal orchestration software*. IRCAM, Paris, first edition, November 2014. [Cited on pages 38 and 41.]

[167] ZAFAR RAFII AND BRYAN PARDO. **Learning to control a reverberator using subjective perceptual descriptors.** In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 285–290, Kobe, Japan, 2009. [Cited on page 38.]

[168] ANDREW TODD SABIN, ZAFAR RAFII, AND BRYAN PARDO. **Weighted-function-based rapid mapping of descriptors to audio processing parameters**. *Journal of the Audio Engineering Society*, **59**(6):419–430, 2011. [Cited on page 38.]

[169] GYÖRGY FAZEKAS. *Semantic Audio Analysis Utilities and Applications*. PhD thesis, Queen Mary University of London, London, UK, 2012. [Cited on pages 38 and 45.]

[170] IVAN EIJI SIMURRA AND JÔNATAS MANZOLLI. **Sound Shizuku composition: a computer-aided composition system for extended music techniques**. *MusMat, Brazilian Journal of Music and Mathematics*, **1**(1):86–101, 2016. [Cited on page 39.]

[171] IVAN EIJI SIMURRA. *Contribuição ao problema da orquestração assistida por computador com suporte de descritores de áudio*. PhD thesis, University of Campinas (Unicamp), São Paulo, Brazil, 2016. [Cited on page 39.]

[172] LÉOPOLD CRESTEL AND PHILIPPE ESLING. **Live Orchestral Piano, a system for real-time orchestral music generation**. In *Proceedings of the 14th Sound and Music Computing Conference (SMC2017)*, pages 434–442, Espoo, Finland, 5–8 July 2017. [Cited on pages 39, 41, and 123.]

[173] STEPHEN MCADAMS. **Timbre as a structuring force in music**. In *Proceedings of Meetings on Acoustics (ICA2013)*, **19**, pages 35–50. Acoustical Society of America, 2013. [Cited on page 39.]

[174] PAUL SMOLENSKY. **Information processing in dynamical systems: Foundations of harmony theory**. In DAVID E. RUMELHART AND JAMES L. MCLELLAND, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 6, pages 194–281. MIT Press, 1986. [Cited on page 39.]

[175] GEOFFREY E HINTON. **A practical guide to training restricted boltzmann machines**. In GRÉGOIRE MONTAVON, GENEVIÈVE B. ORR, AND KLAUS-ROBERT MÜLLER, editors, *Neural networks: Tricks of the trade*, chapter 24, pages 599–619. Springer, 2012. [Cited on page 39.]

[176] JEAN-BAPTISTE BARRIÈRE. *Le Timbre: Métaphore pour la composition*. Christian Bourgois, 1991. [Cited on page 41.]

[177] COLIN G JOHNSON AND ALEX GOUNAROPOULOS. **Timbre interfaces using adjectives and adverbs**. In *Proceedings of the 2006 conference on New interfaces for musical expression*, pages 101–102, Paris, France, 2006. IRCAM—Centre Pompidou. [Cited on page 43.]

[178] PAMELA MCCORDUCK. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. A K Peters, Ltd, Natick, USA, 2nd edition, 2004. [Cited on page 51.]

[179] STUART RUSSELL AND PETER NORVIG. *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, USA, 3rd edition, 2010. [Cited on pages 51, 52, 55, and 59.]

[180] JOHN MCCARTHY, MARVIN L MINSKY, NATHANIEL ROCHESTER, AND CLAUDE E SHANNON. **A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955**. *AI magazine*, **27**(4):12–14, 2006. [Cited on page 51.]

[181] IGOR KONONENKO. **Machine learning for medical diagnosis: history, state of the art and perspective**. *Artificial Intelligence in medicine*, **23**(1):89–109, 2001. [Cited on page 51.]

[182] KESHAV BIMBRAW. **Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology**. In *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, **1**, pages 191–198, Colmar, France, July 2015. IEEE. [Cited on page 51.]

[183] SHANE LEGG AND MARCUS HUTTER. **A collection of definitions of intelligence**. *Frontiers in Artificial Intelligence and applications*, **157**:17–24, 2007. [Cited on page 52.]

[184] LINDA S GOTTFREDSON. **Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography**. *Intelligence*, **24**(1):13–23, 1997. [Cited on page 53.]

[185] ROBERT J STERNBERG. **Contemporary theories of intelligence**. In IRVING B. WEINER, WILLIAM M. REYNOLDS, AND GLORIA E. MILLER, editors, *Handbook of psychology*, **Volume 7, Educational Psychology**, chapter 2, pages 23–45. John Wiley and Sons, Hoboken, USA, 2003. [Cited on page 53.]

[186] JAMES C KAUFMAN, SCOTT BARRY KAUFMAN, AND JONATHAN A PLUCKER. **Contemporary theories of intelligence**. In DANIEL REISBERG, editor, *The oxford handbook of cognitive psychology*, pages 811–822. Oxford University Press, New York, USA, 2013. [Cited on page 53.]

[187] LOUIS LEON THURSTONE. *Primary mental abilities*. University of Chicago Press, Chicago, USA, 1938. [Cited on page 53.]

[188] HOWARD GARDNER. *Frames of mind: The theory of multiple intelligences*. Basic books, New York, USA, 1983. [Cited on pages 53 and 54.]

[189] JOHN B CARROLL. *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press, New York, USA, 1993. [Cited on page 53.]

[190] HOWARD GARDNER. *Multiple intelligences: the theory in practice*. Basic Books, New York, USA, 2011. [Cited on page 54.]

[191] ROBERT J STERNBERG. *Successful intelligence*. Plume, New York, USA, 1997. [Cited on page 54.]

[192] ROBERT J STERNBERG. **Successful intelligence: Finding a balance**. *Trends in cognitive sciences*, **3**(11):436–442, 1999. [Cited on page 54.]

[193] ROBERT J STERNBERG. **The theory of successful intelligence.** *Review of General psychology*, **3**(4):292, 1999. [Cited on page 54.]

[194] ROBERT J STERNBERG. *Beyond IQ: A triarchic theory of human intelligence*. Cambridge University Press, New York, USA, 1985. [Cited on page 54.]

[195] ALAN M TURING. **Computing machinery and intelligence**. *Mind*, **59**:433–460, 1950. [Cited on page 56.]

[196] JON ROWE AND DEREK PARTRIDGE. **Creativity: A survey of AI approaches**. *Artificial Intelligence Review*, **7**(1):43–70, 1993. [Cited on page 57.]

[197] IRÈNE DELIÈGE AND GERAINT A WIGGINS. *Musical creativity: Multidisciplinary research in theory and practice*. Psychology Press, 2006. [Cited on page 57.]

[198] RAY KURZWEIL. *The age of spiritual machines: When computers exceed human intelligence*. Penguin Books, London, UK, 2000. [Cited on page 58.]

[199] RAY KURZWEIL. *The singularity is near: When humans transcend biology*. Penguin Books, London, UK, 2005. [Cited on page 58.]

[200] JUDEA PEARL. *Heuristics: Intelligent search strategies for computer problem solving*. Addison-Wesley Longman Publishing, Boston, USA, 1984. [Cited on page 58.]

[201] KALYANMOY DEB. *Multi-objective optimization using evolutionary algorithms*. John Wiley and Sons, New York, USA, 2001. [Cited on page 59.]

[202] CORINNA CORTES AND VLADIMIR VAPNIK. **Support-vector networks**. *Machine learning*, **20**(3):273–297, 1995. [Cited on pages 59 and 135.]

[203] WARREN S MCCULLOCH AND WALTER PITTS. **A logical calculus of the ideas immanent in nervous activity**. *The bulletin of mathematical biophysics*, **5**(4):115–133, 1943. [Cited on pages 59 and 140.]

[204] KEVIN GURNEY. *An introduction to neural networks*. UCL Press, London, UK, 1997. [Cited on pages 59, 64, and 140.]

[205] NORIS MOHD NOROWI. *An artificial intelligence approach to concatenative sound synthesis*. PhD thesis, University of Plymouth, Plymouth, UK, 2013. [Cited on page 60.]

[206] DIEMO SCHWARZ. *Data-driven concatenative sound synthesis*. PhD thesis, Université Paris 6 (Pierre et Marie Curie), Paris, France, 2004. [Cited on page 60.]

[207] EDUARDO RECK MIRANDA. **Striking the right note with ARTIST: an AI-based synthesiser**. *Recherches et applications en informatique musicale*, pages 227–239, 1998. [Cited on page 60.]

[208] ANDREW HORNER, JAMES BEAUCHAMP, AND LIPPOLD HAKEN. **Machine tongues XVI: Genetic algorithms and their application to FM matching synthesis**. *Computer Music Journal*, **17**(4):17–29, 1993. [Cited on page 61.]

[209] ANDREW OWENS, PHILLIP ISOLA, JOSH MCDERMOTT, ANTONIO TORRALBA, EDWARD H ADELSON, AND WILLIAM T FREEMAN. **Visually indicated sounds**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016. [Cited on page 61.]

[210] GEOFFREY E HINTON, SIMON OSINDERO, AND YEE-WHYE TEH. **A fast learning algorithm for deep belief nets**. *Neural computation*, **18**(7):1527–1554, 2006. [Cited on page 61.]

[211] YOSHUA BENGIO ET AL. **Learning deep architectures for AI**. *Foundations and trends® in Machine Learning*, **2**(1):1–127, 2009. [Cited on page 61.]

[212] AARON VAN DEN OORD, SANDER DIELEMAN, HEIGA ZEN, KAREN SIMONYAN, ORIOL VINYALS, ALEX GRAVES, NAL KALCHBRENNER, ANDREW SENIOR, AND KORAY KAVUKCUOGLU. **Wavenet: A generative model for raw audio**. *arXiv preprint arXiv:1609.03499*, 2016. [Cited on page 61.]

[213] JESSE ENGEL, CINJON RESNICK, ADAM ROBERTS, SANDER DIELEMAN, DOUGLAS ECK, KAREN SIMONYAN, AND MOHAMMAD NOROUZI. **Neural audio synthesis of musical notes with wavenet autoencoders**. *arXiv preprint arXiv:1704.01279*, 2017. [Cited on page 61.]

[214] DAVID BESSELL. **Dynamic convolution modeling, a hybrid synthesis strategy**. *Computer Music Journal*, **37**(1):44–51, 2013. [Cited on page 62.]

[215] DAN DUGAN. **Automatic microphone mixing**. *Journal of the Audio Engineering Society*, **23**(6):442–449, 1975. [Cited on page 62.]

[216] MARK BROZIER CARTWRIGHT AND BRYAN PARDO. **Social-EQ: Crowdsourcing an equalization descriptor map**. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 395–400, Curitiba, Brazil, 2013. [Cited on page 62.]

[217] BRECHT DE MAN AND JOSHUA D REISS. **A knowledge-engineered autonomous mixing system**. In *135th Audio Engineering Society Convention*, New York, USA, 2013. Audio Engineering Society. [Cited on page 63.]

[218] BRECHT DE MAN AND JOSHUA D REISS. **A semantic approach to autonomous mixing**. *Journal on the Art of Record Production (JARP)*, **08**, December 2013. [Cited on page 63.]

[219] IVER JORDAL, ØYVIND BRANDTSEGG, AND GUNNAR TUFTE. **Evolving neural networks for cross-adaptive audio effects**. In *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP)*, London, UK, 2016. [Cited on page 63.]

[220] ALEX WILSON AND BRUNO FAZENDA. **An evolutionary computation approach to intelligent music production informed by experimentally gathered domain knowledge**. In *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP)*, London, UK, 2016. [Cited on page 63.]

[221] JOSE D FERNÁNDEZ AND FRANCISCO VICO. **AI methods in algorithmic composition: A comprehensive survey**. *Journal of Artificial Intelligence Research*, **48**:513–582, 2013. [Cited on page 64.]

[222] LEJAREN HILLER AND LEONARD MAXWELL ISAACSON. *Experimental music: composition with an electronic computer*. McGraw-Hill, New-York, 1959. [Cited on pages 64 and 66.]

[223] RONALD A HOWARD. *Dynamic probabilistic systems: Markov models*. Dover Publications, New York, USA, 2007. [Cited on pages 64 and 176.]

[224] JAMES ANDERSON MOORER. **Music and computer composition**. *Communications of the ACM*, **15**(2):104–113, 1972. [Cited on page 64.]

[225] SEVER TIPEI. **MP1: a computer program for music composition**. In *2nd Annual Music Computation Conference*, pages 68–82, Urbana, USA, 1975. [Cited on page 64.]

[226] FRANÇOIS PACHET. **Interacting with a musical learning system: The continuator**. In *International Conference on Music and Artificial Intelligence*, pages 103–108. Springer, 2002. [Cited on page 64.]

[227] STEPHEN DAVISMOON AND JOHN ECCLES. **Combining musical constraints with Markov transition probabilities to improve the generation of creative musical structures**. In *European Conference on the Applications of Evolutionary Computation*, pages 361–370. Springer, 2010. [Cited on page 64.]

[228] PETER M TODD. **A connectionist approach to algorithmic composition**. *Computer Music Journal*, **13**(4):27–43, 1989. [Cited on page 65.]

[229] NAOKI SHIBATA. **A neural network-based method for chord/note scale association with melodies**. *NEC research and development*, **32**(3):453–459, 1991. [Cited on page 65.]

[230] PETRI TOIVIAINEN. **Modeling the target-note technique of bebop-style jazz improvisation: An artificial neural network approach**. *Music Perception: An Interdisciplinary Journal*, **12**(4):399–413, 1995. [Cited on page 65.]

[231] ALAN DORIN. **Boolean networks for the generation of rhythmic structure**. In *Proceedings of the Australian Computer Music Conference*, **38**, page 45, 2000. [Cited on page 65.]

[232] DAVID A FERRUCCI. **Introduction to "this is watson"**. *IBM Journal of Research and Development*, **56**(3.4):1–15, 2012. [Cited on page 65.]

[233] DAVID COPE. *Experiments in musical intelligence*. AR Editions, Madison, USA, 1996. [Cited on page 67.]

[234] DAVID COPE. *Computer models of musical creativity*. MIT Press, Cambridge, USA, 2005. [Cited on page 67.]

[235] CARLOS SÁNCHEZ QUINTANA, FRANCISCO MORENO ARCAS, DAVID ALBARRACÍN MOLINA, JOSÉ DAVID FERNÁNDEZ RODRIGUEZ, AND FRANCISCO J VICO. **Melomics: A case-study of AI in Spain**. *AI Magazine*, **34**(3):99–103, 2013. [Cited on page 67.]

[236] GUSTAVO DIAZ-JEREZ. **Composing with Melomics: Delving into the computational world for musical inspiration**. *Leonardo Music Journal*, **21**:13–14, 2011. [Cited on page 67.]

[237] FRANÇOIS PACHET, PIERRE ROY, AND FIAMMETTA GHEDINI. **Creativity through style manipulation: the flow machines project**. In *The 2013 Marconi Institute for Creativity Conference*, Bologna, Italy, 2013. [Cited on page 68.]

[238] FIAMMETTA GHEDINI, FRANÇOIS PACHET, AND PIERRE ROY. **Creating music and texts with flow machines**. In *Multidisciplinary Contributions to the Science of Creative Thinking*, pages 325–343. Springer, 2016. [Cited on page 68.]

[239] FRANÇOIS PACHET. **A joyful ode to automatic orchestration**. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **8**(2):18, 2016. [Cited on page 68.]

[240] GAËTAN HADJERES, FRANÇOIS PACHET, AND FRANK NIELSEN. **DeepBach: a steerable model for Bach chorales generation**. *arXiv preprint arXiv:1612.01010v2*, 2017. [Cited on page 68.]

[241] JEAN-PIERRE BRIOT, GAËTAN HADJERES, AND FRANÇOIS PACHET. **Deep learning techniques for music generation - a survey**. *arXiv preprint arXiv:1709.01620*, 2017. [Cited on page 71.]

[242] ERNST TERHARDT. **Frequency analysis and periodicity detection in the sensations of roughness and periodicity pitch**. In REINIER PLOMP AND GUIDO F SMOORENBURG, editors, *Frequency analysis and periodicity detection in hearing*, pages 278–290. Sijthoff Leiden, The Netherlands, 1970. [Cited on page 79.]

[243] ILSE BERNADETTE LABUSCHAGNE AND VALTER CIOCCA. **The perception of breathiness: Acoustic correlates and the influence of methodological factors**. *Acoustical Science and Technology*, **37**(5):191–201, 2016. [Cited on pages 81 and 106.]

[244] DEIRDRE BRIGID BOLGER. *Computational models of musical timbre and the analysis of its structure in melody*. PhD thesis, University of Limerick, 2004. [Cited on pages 81 and 107.]

[245] HARVEY FLETCHER. **Auditory patterns**. *Reviews of modern physics*, **12**(1):47, 1940. [Cited on page 82.]

[246] VON W AURES. **A procedure for calculating auditory roughness**. *Acustica*, **58**(5):268–281, 1985. [Cited on pages 82 and 107.]

[247] RUSS ETHINGTON AND BILL PUNCH. **SeaWave: A system for musical timbre description**. *Computer Music Journal*, pages 30–39, 1994. [Cited on pages 82 and 107.]

[248] OLIVIER LARTILLOT, PETRI TOIVIAINEN, AND TUOMAS EEROLA. **A matlab toolbox for music information retrieval**. In *Data analysis, machine learning and applications*, pages 261–268. Springer, 2008. [Cited on page 85.]

[249] PATRICIA KEATING, MARC GARELLEK, AND JODY KREIMAN. **Acoustic properties of different kinds of creaky voice**. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, 2015. [Cited on page 86.]

[250] PATRIK N JUSLIN. **Cue utilization in communication of emotion in music performance: Relating performance to perception.** *Journal of Experimental Psychology: Human perception and performance*, **26**(6):1797, 2000. [Cited on page 86.]

[251] PETRI LAUKKA, PATRIK JUSLIN, AND ROBERTO BRESIN. **A dimensional approach to vocal expression of emotion**. *Cognition and Emotion*, **19**(5):633–653, 2005. [Cited on page 86.]

[252] JAMES B. MACQUEEN. **Some methods for classification and analysis of multivariate observations**. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, **1**, pages 281–297. Oakland, CA, USA, 1967. [Cited on page 123.]

[253] FABIAN PEDREGOSA, GAËL VAROQUAUX, ALEXANDRE GRAMFORT, VINCENT MICHEL, BERTRAND THIRION, OLIVIER GRISEL, MATHIEU BLONDEL, PETER PRETTENHOFER, RON WEISS, VINCENT DUBOURG, ET AL. **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, **12**(Oct):2825–2830, 2011. [Cited on page 125.]

[254] DAVID ARTHUR AND SERGEI VASSILVITSKII. **k-means++: The advantages of careful seeding**. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. [Cited on page 126.]

[255] EDGAR OSUNA, ROBERT FREUND, AND FEDERICO GIROSIT. **Training support vector machines: an application to face detection**. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136. IEEE, 1997. [Cited on page 135.]

[256] VINCENT WAN AND WILLIAM M CAMPBELL. **Support vector machines for speaker verification and identification**. In *Proceedings of IEEE Workshop on Neural Networks for Signal Processing X*, **2**, pages 775–784. IEEE, 2000. [Cited on page 135.]

[257] SCIKIT-LEARN. **API Reference**. http://scikit-learn.org/0.18/modules/classes.html [Accessed: 09/12/2017]. [Cited on page 136.]

[258] SEPP HOCHREITER AND JÜRGEN SCHMIDHUBER. **Long short-term memory**. *Neural computation*, **9**(8):1735–1780, 1997. [Cited on page 140.]

[259] ALEX KRIZHEVSKY, ILYA SUTSKEVER, AND GEOFFREY E HINTON. **Imagenet classification with deep convolutional neural networks**. In *Advances in neural information processing systems*, **25**, pages 1097–1105, 2012. [Cited on page 140.]

[260] DIEDERIK P KINGMA AND JIMMY BA. **Adam: A method for stochastic optimization**. *arXiv preprint arXiv:1412.6980*, 2014. [Cited on page 141.]

[261] RICHARD S SUTTON AND ANDREW G BARTO. *Reinforcement learning: An introduction*, **1**. MIT Press, Cambridge, 1998. [Cited on page 144.]

[262] ROGER FLETCHER. *Practical methods of optimization*. John Wiley and Sons, 2nd edition, 2000. [Cited on page 171.]

[263] JEAN-MICHAEL CELERIER, PASCAL BALTAZAR, CLEMENT BOSSUT, NICOLAS VUAILLE, JEAN-MICHEL COUTURIER, AND MYRIAM DESAINTE-CATHERINE. **OSSIA: Towards a unified interface for scoring time and interaction**. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation - TENOR2015*, pages 81–90, Paris, France, 2015. Institut de Recherche en Musicologie. [Cited on page 215.]

[264] ANTOINE ALLOMBERT, GERARD ASSAYAG, AND MYRIAM DESAINTE-CATHERINE. **A system of interactive scores based on Petri nets**. In *Proceedings of the 4th Sound and Music Computing Conference (SMC07)*, pages 158–165, Lefkada, Greece, 2007. [Cited on page 215.]